Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)
Кафедра Штучного інтелекту
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Анонімізація великих даних в інтелектуальних медичних системах

(тема)

Виконав:
студент 2 курсу, групи СШМ-22-3
Воронова Д. С.
(прізвище, ініціали)

Спеціальність 122 Комп'ютерні науки
(код і повна назва спеціальності)
Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)
Освітня програма Системи штучного інтелекту
(повна назва спеціалізації)
Керівник доц. Чала Л.Е.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ В.О. Філатов
(підпис) (прізвище, ініціали)

2024 р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)

Кафедра Штучного інтелекту
(повна назва)

Рівень вищої освіти другий (магістерський)

Спеціальність 122 Комп'ютерні науки
(код і повна назва)

Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма Системи штучного інтелекту
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«_____» _____ 20 ___ р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

студентові Вороновій Дар'ї Сергіївні
(прізвище, ім'я, по батькові)

1. Тема роботи Анонімізація великих даних в інтелектуальних медичних системах

затверджена наказом університету від 1 квітня 20 24 р. № 260Ст

2. Термін подання студентом роботи до екзаменаційної комісії 04 червня 20 24 р.

3. Вихідні дані до роботи науково-технічні публікації, дані інтернет-джерел та відомих науукових проєктів щодо анонімізації даних, Python Документація, набори даних з реальними показниками в медичних системах

4. Перелік питань, що потрібно опрацювати в роботі
1) Аналіз предметної галузі
2) Огляд існуючих рішень в контексті анонімізації даних
3) Методи та алгоритми анонімізації даних
4) Розроблення моделей для анонімізації шляхом створення синтетичних даних
5) Оцінка результатів експерименту

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Отримання завдання на кваліфікаційну роботу	01.04.2024	виконано
2	Аналіз поставленого завдання	05.04.2024	виконано
3	Огляд інформаційних джерел із заданої тематики	14.04.2024	виконано
4	Огляд існуючих рішень та формулювання вимог	19.04.2024	виконано
5	Пошук та аналіз набору даних	23.04.2024	виконано
6	Огляд засобів розробки	26.04.2024	виконано
7	Навчання моделей	28.04.2024	виконано
8	Розробка та тестування програмного продукту	01.05.2024	виконано
9	Написання пояснювальної записки	14.05.2024	виконано
10	Перевірка на академічний плагіат	15.05.2024	виконано
11	Нормоконтроль	16.05.2024	виконано
12	Підготовка презентації та доповіді	31.05.2024	виконано
13	Попередній захист	01.06.2024	виконано
14	Рецензування	02.06.2024	виконано
15	Захист перед екзаменаційною комісією	04.06.2024	виконано

Дата видачі завдання 1 квітня 20 24 р.

Студент _____
(підпис)

Керівник роботи _____ доц. Чала Л.Е.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка: 93 с., 27 рис., 16 табл., 2 дод., 30 джерел.

АНОНІМІЗАЦІЯ ДАНИХ, ІНТЕЛЕКТУАЛЬНІ МЕДИЧНІ СИСТЕМИ, МАШИННЕ НАВЧАННЯ, СИНТЕТИЧНІ ДАНІ.

Об'єкт дослідження – інтелектуальні медичні системи.

Предмет дослідження – застосування методів анонімізації даних у медичних системах.

Мета роботи – розробка та дослідження ефективності методів штучного інтелекту для анонімізації медичних даних. Роботу присвячено вдосконаленню процесу захисту медичних даних забезпечення конфіденційності пацієнтів та забезпечення відповідності нормативно-правовим вимогам у сфері захисту персональних даних у медичних системах.

Методи дослідження – аналіз існуючих методів анонімізації багатовимірних даних у медичній сфері. Цей аналіз передбачає оцінку ефективності, точності та відповідності вимогам конфіденційності кожного методу. Під час аналізу враховувались особливості медичних даних та вимоги до їх обробки. Від цього залежать вибір оптимальних стратегій анонімізації, що найкраще відповідають контексту медичної інформації та гарантують збереження якості даних. Передбачається реалізація обраної стратегії анонімізації та її впровадження у інтелектуальні медичні системи. Це включає в себе розробку відповідних алгоритмів та створення програмного коду.

Дослідження виконано на основі матеріалів та завдяки доступу до експертного середовища, який є результатом проєкту WARN «Академічна протидія гібридним загрозам» (610133-EPP-1-2019-1-FI-EPPKA2-CBHE-JP), що співфінансується програмою Erasmus+ Європейського Союзу.

ABSTRACT

Master's thesis contains: 93 p., 27 fig., 16 tabl., 2 ann., 30 references.

DATA ANONYMISATION, SMART HEALTHCARE SYSTEMS, MACHINE LEARNING, SYNTHETIC DATA.

The object of research is smart healthcare systems.

The subject of research is an application of methods of anonymisation of big medical data.

The purpose of the work is to develop and study the effectiveness of artificial intelligence methods for anonymising medical data. The work is dedicated to improving the process of healthcare data protection to ensure patient confidentiality and compliance with legal requirements in the field of personal data protection in medical systems.

The research methods are analysis of existing methods of anonymising big data in the healthcare field. This analysis involves assessing the efficiency, accuracy and compliance with privacy requirements of each method. The analysis accounts for the specifics of medical data and the requirements for its processing, ensuring compliance with legal and ethical standard. This determines the choice of optimal anonymisation strategies that best suit the context of medical information and guarantee data quality. It is planned to implement the selected anonymisation strategy and its implementation in intelligent healthcare systems. This includes the development of appropriate algorithms and the creation of programme code.

The research is based on materials and thanks to access to the expert environment resulting from the WARN project «Academic Response to Hybrid Threats» (610133-EPP-1-2019-1-FI-EPPKA2-CBHE-JP), co-funded by the Erasmus+ Programme of the European Union.

ЗМІСТ

Перелік умовних позначень, символів, одиниць, скорочень і термінів 8
1 Аналіз предметної області та постановка задачі 11
1.1 Аналіз предметної області 11
1.2. Регламент захисту даних GDPR.....19
1.3 Загрози даним 20
1.4 Постановка задач дослідження 22
2 Методи анонімізації даних 24
2.1 Анонімізація та конфіденційність даних 24
2.2 Методи анонімізації та псевдоанонімізації 28
2.2.1 Традиційні методи анонімізації 29
2.2.2 Просунуті методи анонімізації 33
2.3 Атаки на дані.....35
3 Анонімізація медичних даних. огляд рішень.....44
3.1 Реляційний модель представлення даних 44
3.2 Графічні дані про стан здоров'я 46
3.3 Генерування синтетичних даних.....49
3.4 Генеративні змагальні мережі.....50
3.5 Трансформери.....52
3.6 Порівняння GAN та трансофрмерів 54
3.7 medGAN 55
4 Програмна реалізація генерування синтетичних даних.....57
4.1 Огляд засобів розробки.....57
4.2 Аналіз наборів даних 58
4.3 Покращені моделі 65
4.4 Опис експерименту.....66
4.5 Оцінка експерименту 68
4.6 Результати експерименту.....71
Висновки 86

Перелік джерел посилання 88
Додаток А Сертифікати про участь у конференціях з теми кваліфікаційної роботи.....91
Додаток Б Відомість кваліфікаційної роботи.....93

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

ЕОПП – електронний охоронний протокол пацієнта;

ЕФР – емпірична функція розподілу;

МСХ – міжнародна мтатистична класифікація хвороб;

РНОКПП – реєстраційний номер облікової картки платника податків;

AUC – Area Under the Curve – площа під кривою;

CAHD – Correlation-Aware High-dimensional Data Anonymization – анонімізація великих даних з урахуванням кореляції;

CC – Correlation Coefficients – коефіцієнт кореляції;

DID – Direct Identifier – прямий ідентифікатор;

DL – Deep Learning – глибоке навчання;

DP – Data Pseudonymization – псевдоанонімізація даних;

EMD – Earth Mover's Distance;

FFNN – Feedforward Neural Network – нейронна мережа з прямим розповсюдженням;

GAN – Generative Adversarial Network – генеративні змагальні мережі;

GDPR – General Data Protection Regulation – загальний регламент про захист даних;

HE – Homomorphic Encryption – гомоморфне шифрування;

KL – Kullback-Leibler Divergence – розходження Кульбака-Лейблера;

LKC – Latent Knowledge Consistency — узгодження латентних знань;

NLP – Natural Language Processing – обробка природної мови;

QID – Quasi Identifier – квазіідентифікатор;

RMSE – Root Mean Square Errors – середньоквадратичні помилки;

SA – Sensitive Attribute – чутливий атрибут;

SVM – Support Vector Machine – метод опорних векторів;

UAA – Utility-Aware Anonymization – анонімізація з урахуванням корисності.

ВСТУП

Сучасна медицина переживає період надзвичайного розвитку завдяки поєднанню двох потужних факторів: великих медичних даних і штучного інтелекту. Великі медичні дані відкривають нові можливості для більш точних діагнозів та персоналізованого лікування, тоді як штучний інтелект дозволяє автоматизувати прогнозування захворювань, оптимізувати процес лікування та підтримувати прийняття рішень.

Оскільки сучасна медична інфраструктура все більше стає цифровою та підключеною до мережі, вона все частіше стає об’єктом гібридних загроз. Медичні системи стають привабливою мішенню для зловмисників з різними мотивами: від фінансових вигод до геополітичних цілей. Атаки на медичні системи, як правило, спрямовані на викрадення конфіденційних медичних даних для вимагання викупу, порушення нормальної роботи медичних пристроїв, і навіть зміну медичних записів з метою спотворення діагнозів або лікування пацієнтів. Вони можуть також спричинити загрозу життю і здоров'ю пацієнтів, руйнуючи довіру до медичних послуг та інфраструктури загалом. Для вирішення цих питань використовуються різні підходи, зокрема матеріали з вивчення та протидії гібридним загрозам, отримані в рамках роботи WARN-середовища (проєкт WARN «Академічна протидія гібридним загрозам» 610133-EPP-1-2019-1-FI-EPPKA2-CBHE-JP).

Зберігання та обробка медичних даних відповідно до встановлених стандартів конфіденційності та безпеки є критично важливою. Етичні стандарти визначають правила поведінки та взаємодії в медичній сфері, гарантуючи захист конфіденційності та прав пацієнтів, а також етичне використання їхніх особистих даних. Хакерські атаки на медичні системи призводять до витоку конфіденційної інформації, яку зловмисники можуть використати для різних видів шахрайства та блокування роботи важливих медичних систем, створюючи пряму загрозу безпеці пацієнтів. Один зі способів захисту медичних даних – анонімізація, іншими словами,

видалення ідентифікаторів особи, а також узагальнення будь-яких ознак, які можуть бути використані для встановлення зв'язку з конкретною людиною.

Метою кваліфікаційної роботи є розроблення методу анонімізації великих даних з використанням методів машинного навчання, які доцільно використовувати, коли необхідно поєднати конфіденційність даних із збереженням структури та статистичних властивостей даних для ефективного навчання інтелектуальних моделей. Для програмної реалізації в даному проекті використовуєтья Python – високорівнева мова програмування загального призначення, яка стала основою для розробки великої кількості бібліотек для машинного навчання і обробки даних.

Подяка (acknowledgment): Дослідження виконано на основі матеріалів та завдяки доступу до експертного WARN-середовища, який є результатом проєкту WARN «Академічна протидія гібридним загрозам» (610133-EPP-1-2019-1-FI-EPPKA2-CBHE-JP «Academic Response to Hybrid Threats»), що співфінансується програмою Erasmus+ Європейського Союзу.

Застереження (disclaimer): «Підтримка Європейською Комісією випуску цієї публікації не означає схвалення змісту, який відображає лише погляди авторів, і Комісія не несе відповідальності за будь-які використання інформації, що міститься в ньому» («The European Commission's support for the production of this publication does not constitute an endorsement of the contents, which reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein»).

1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ТА ПОСТАНОВКА ЗАДАЧІ

1.1 Аналіз предметної області

Здоров’я – основа нашого життя. Без цього важливого фактора неможливе повноцінне існування людини, духовний, психологічний та емоційний стан. Здоров’я дає можливість жити повним життям – вчитися, займатися спортом, мандрувати, розвиватися як особистість. Здорова людина – щаслива людина. Важливо піклуватися про свій стан аби відчути життя на повну, насолодитись й спробувати все в ньому. Можливість зберігати та порівнювати інформацію про стан свого організму – важливий крок до піклування про себе і підтримки комфортного рівню свого життя. Багато людей не приділяють достатньої уваги стану свого організму, а потім стикаються з різними хворобами. Мало хто приносить із собою до лікаря свої аналізи за декілька останніх років, але, наприклад, стан крові в період попереднього захворювання може показати на якій стадії і як протікає хвороба цього разу.

До того ж, може статися, що окремо взяті аналізи будуть у нормі, але у порівнянні з попередніми даними буде видно, що деякі показники підвищились або знизились, що не є нормою. Це може свідчити про порушення функцій якогось органа або системи організму. Маючи можливість відстежувати показники аналізів у динаміці – безперечна перевага для пацієнта. А лікарі завдяки тому, що бачать динаміку, можуть призначити правильне лікування та при необхідності корегувати його у потрібний момент. Це дасть змогу попередити деякі захворювання та/або своєчасно почати лікування або дати організму відпочити. Якщо с таким додатком будуть працювати лабораторії і результати тестів будуть надходити безпосередньо у кабінет пацієнта і інтегруватись в дані, то цінність такого додатка як для лікаря, так і пацієнта дуже велика. До того ж люди із хронічними захворюваннями зможуть відстежувати динаміку

показників у таблицях або діаграмах і, у разі поліпшення стану, буде з'являтися мотивація лікуватись далі та не втрачати сподівання на одужання.

Співпраця лікаря та пацієнта повинна бути зручною та наглядною. Лікар зможе краще оцінити ситуацію та стан хворого, маючи на руках висновки попередніх хвороб та маючи змогу порівняти показники аналізів, або, для економії часу, виводити необхідні дані у таблицю за обраний період.

Сучасний світ розвивається душе швидко, багато людей веде дуже активний спосіб життя. Все більше і більше людей приділяє багато уваги онлайн життю. Кожного дня зв'являються все більше і більше різноманітних додатків для різних сфер нашого життя, серед них є і пов'язані з медициною.

На даний момент існує велика кількість систем орієнтованих на лікарів або студентів медичних навчальних закладів. Дуже багато сучасних додатків пов'язані з довідковою літературою, описом захворювань та лікарських препаратів. Деякі з додатків мають вузьку спеціалізацію, наприклад, за видами лікарської справи, тож є необхідність встановлення одразу декількох. Інший напрямок розробки медичних додатків – скед'юлери для полегшення запам'ятовування та відслідковування прийому лікарських препаратів.

Нижче наведено декілька додатків, які використовуються для поєднання та полегшення використавння медичних даних в умовах діджиталізації:

CareZone [1] додаток дає можливість не забувати про своєчасний прийом ліків, дає змогу створити власний календар, дивитися які лікі приймалися раніше та порівнювати своє лікування з попередніми хворобами, або робити помітки, які ліки були більш ефективними (рисунки 1.1 та 1.2).

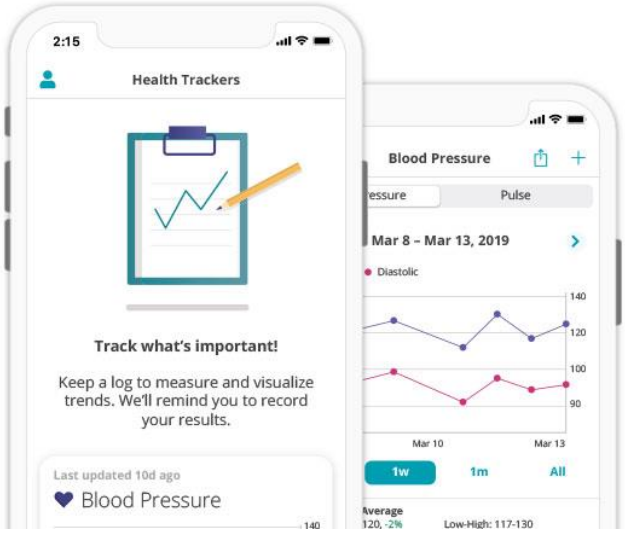

Рисунок 1.1 – Приклад діаграми відстеження кров'яного тиску

Рисунок 1.2 – Приклад занесення необхідних ліків, з можливістю відсканування

Переваги:

– можливість структурувати прийом ліків;
– універсальний додаток, не має обмежень у застосуванні;
– знижує можливість пропустити прийом ліків;

– зберігає історію лікування, що зручно, бо можна звертатися до цієї інформацію у будь який час.

Недоліки:

– немає можливості вносити та порівнювати дані стосовно аналізів;

– немає статистичних даних про стан здоров’я.

Doctor Booster (Doctor Online) – додаток, який допомагає відстежити перебіг захворювання у пацієнта (рисунки 1.3 та 1.4).

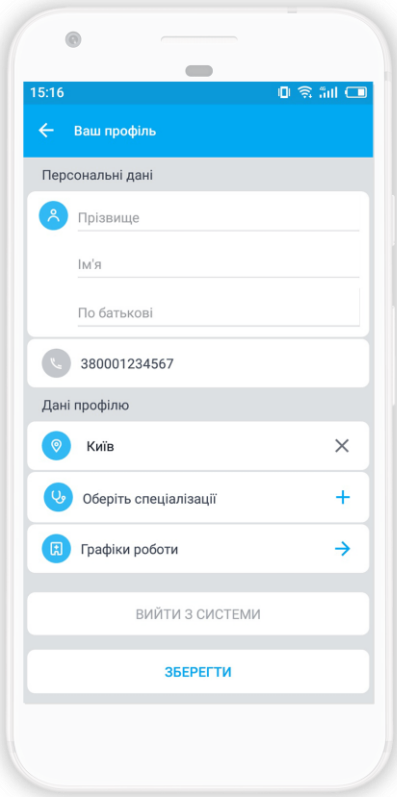

Рисунок 1.3 – Форма реєстрації лікаря

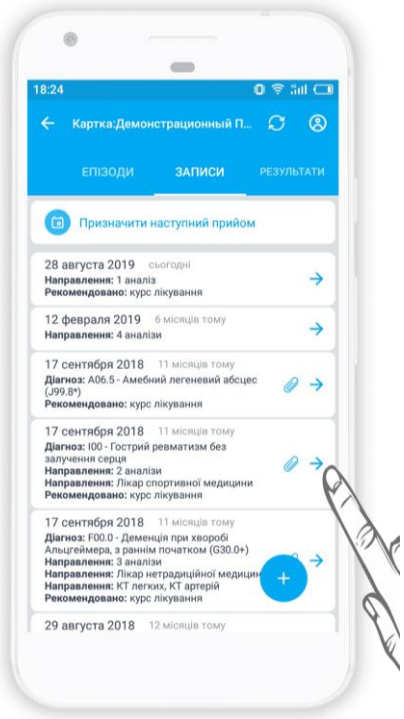

Рисунок 1.4 – Огляд даних про пацієнта, вибір даних для експорту

Переваги:

– дає змогу бачити загальний стан та динаміку захворювання;

– є можливість надавати електронні направлення на аналізи;

– можливість додавати та створювати графік прийому ліків.

Недоліки:

– орієнтовний на лікарів, а не на пацієнтів;

– немає можливості вносити та порівнювати дані стосовно аналізів.

Pregnancy [2] має широкий спектр можливостей стосовно ведення вагітності (рисунки 1.5 та 1.6).

Переваги:

– є можливість додавати та зберігати інформацію, у тому числі скрінінг-УЗД.

Недоліки:

– немає можливості порівнювати дані про аналізи;

– має вузьку спеціалізацію.

Рисунок 1.5 – Сторінка відстеження ваги вагітної

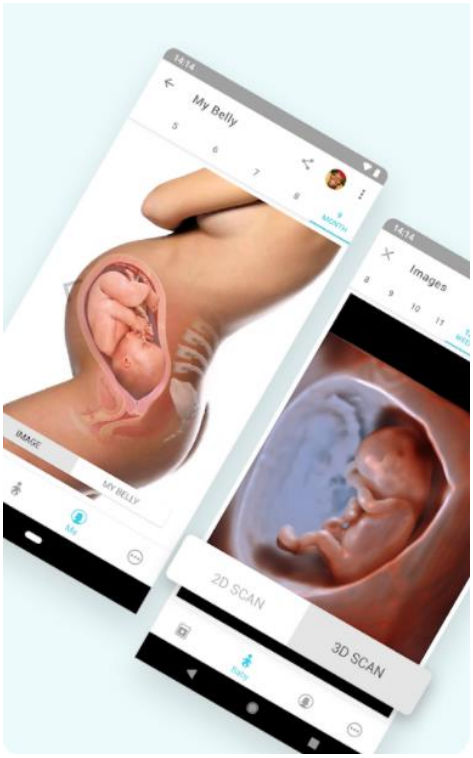

Рисунок 1.6 – Сторінка зі збереженими скринінг-УЗД

Діапазон референтних значень в різних лабораторіях може відрізнятись. Розроблений у попередній роботі [3] додаток позиціонує себе як універсальний помічник, що підходить для будь-яких лабораторії. Дані нормалізуються до одного універсального показника, аби всі дані знаходились в одному діапазоні, а також одиницях вимірювання, тому при порівнянні та побудові діаграм на основі цих даних, не виникає помилок у коректності відображення або питань щодо великих різниць у показниках.

Додаток має можливість про нагадування про необхідність пройти ти чи інші аналізі для порівняння даних, або для відстеження перебігу захворювання.

Види аналізів, які доступні для завантаження в запропонований додаток [4]:

– загальноклінічні дослідження;
– біохімічні аналізи;
– бактеріологічні дослідження;
– імунологічні дослідження;
– цитологічні дослідження;
– паразитарні дослідження;
– гормональні дослідження;
– онкологічні маркери;
– молекулярно-біологічні дослідження (ПЛР-діагностика), тощо.

Для зручної роботи лікаря запропонований додаток можна інтегрувати в роботу HELSI (рисунок 1.7) або інших медичних застосунків.

Чому першочергова мета – інтеграція з HELSI? Ця система зручна через ряд причин, а саме [5]:

– можливість легко знайти та обрати свого лікаря;
– швидкий запис на прийом on-line себе та членів своєї родини;

– доступ до своєї ЕМК;
– миттєві результати аналізів та діагностики в кабінеті пацієнта;
– доступ до призначень лікаря та плану лікування.

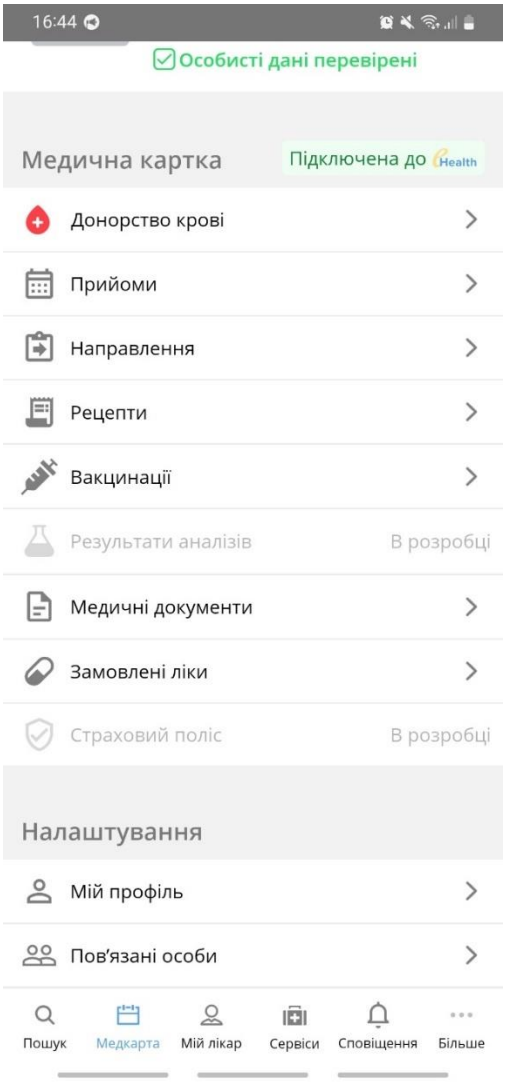

Рисунок 1.7 – Інтерфейс мобільного додатку HELSI

В останні десятиліття в медичній сфері спостерігається значний розвиток інтелектуальних систем, які використовують аналіз даних для розпізнавання захворювань, передбачення медичних подій та оптимізації лікувальних процесів. Однак збільшення обсягів медичних даних із різних джерел (лікарські картки, лабораторні дослідження тощо) створює серйозні

виклики з погляду конфіденційності та приватності пацієнтів. З метою захисту цих даних передбачається використання технологій анонімізації.

Проблема полягає в тому, що анонімізація багатовимірних медичних даних, тобто процес заміни ідентифікаторів (наприклад, імен, номерів страхових полісів) на неідентифікуючі або псевдоідентифікуючі дані, має бути виконана таким чином, щоб зберегти корисну інформацію для подальшого аналізу та досліджень, а також забезпечити недоступність особистої інформації пацієнтів перед неповідомленими особами.

1.2. Регламент захисту даних GDPR

Регламент загального застосування про захист даних (GDPR) є ключовим законодавством Європейського Союзу, прийнятим з метою забезпечення захисту особистих даних громадян ЄС. Цей регламент має велике значення для бізнесу, оскільки встановлює обов'язки та вимоги щодо обробки даних, включаючи їх зберігання, передачу та захист.

Одним із ключових принципів GDPR є забезпечення прав осіб на захист їхніх даних. Згідно з регламентом, особи мають право на доступ до своїх особистих даних, їх виправлення, видалення та перенесення до іншої організації. Ці права забезпечують контроль осіб над їхніми даними та сприяють збереженню їхньої приватності.

Окрім цього, GDPR встановлює вимоги щодо захисту даних, зокрема шляхом впровадження технічних та організаційних заходів, спрямованих на запобігання несанкціонованому доступу, втраті, зміні чи поширенню даних. Ці заходи включають у себе шифрування, аудит доступу до даних та регулярні перевірки безпеки [6].

Узагальнюючи, GDPR встановлює важливі стандарти та вимоги щодо захисту особистих даних, що сприяє забезпеченню прозорості, контролю та безпеки в обробці даних в Європейському Союзі. Виконання цих вимог

відображає відповідальний підхід до обробки особистих даних та сприяє зміцненню довіри споживачів.

Отже, відповідно до GDPR необхідно:

- збирати особисті дані тільки з дозволу суб'єкта;
- використовувати та обробляти їх відповідно до визначених цілей;
- після досягнення цілей – знищувати дані;
- вилучати та знищувати дані за запитом їх власника;
- забезпечувати безпеку зберігання даних;
- не передавати дані без згоди суб'єкта.

1.3 Загрози даним

У відповідності з правилами GDPR, не існують конкретних вимог щодо методів та порядку захисту даних. Кожна організація має право вибирати заходи захисту самостійно. Одним з ключових аспектів GDPR є анонімізація особистих даних при їх передачі третім сторонам або опублікуванні.

Процес анонімізації, або редагування даних, полягає у вилученні або приховуванні особистої інформації з метою подальшого використання даних. Хоча на перший погляд може здатися, що проблему можна вирішити шляхом простого видалення стовпців з особистою інформацією з датасету, насправді це складніше. Ілюстрацією цього може служити інцидент, що стався в штаті Массачусетс [7]. Комісія з групового страхування (GIC), державне агентство, що надавало медичне страхування державним службовцям, оприлюднила набір даних про візити до лікаря державних службовців. Намагаючись захистити конфіденційність, GIC видалила всі імена, адреси та інші особисті ідентифікатори. Однак, вчений Суїні [7] зміг ідентифікувати медичну інформацію губернатора Вільяма Уельда, використовуючи дані перепису населення 1990 року. Через десятиліття,

Голле [8] провів перерахунок статистики, використовуючи дані перепису населення 2000 року. Застосовуючи ту ж комбінацію атрибутів QID, яка включала дату народження, стать та поштовий індекс, Голлу вдалося однозначно ідентифікувати 63% населення США.

Подібні проблеми виникали й у інших контекстах. Наприклад, в 2006 році сервіс онлайн-прокату фільмів Netflix запропонував приз у мільярд доларів за покращення їхньої системи рекомендацій [9]. Netflix опублікував набір даних, який містив рейтинги фільмів, оцінені майже 500000 підписниками, разом з назвами фільмів та їх рейтингами. З метою захисту приватності користувачів, всі особисті ідентифікатори були вилучені з набору даних і замінені випадковими ідентифікаторами. Проте в дослідженні [10] виявлено, що знання дат (±2 тижні) шести рейтингів фільмів дозволяє ідентифікувати 99% осіб у наборі даних. Крім того, знання лише двох фільмів з датами рейтингу (±3 дні) дозволяє ідентифікувати 68% підписників з цього набору. Автори використали загальнодоступний, неанонімізований набір даних з оглядів фільмів на веб-сайті Internet Movie Database (IMDb) для своєї атаки.

Компанія America Online (AOL) оприлюднила набір даних, що містив 20 мільйонів пошукових запитів 650 000 користувачів з метою залучення уваги дослідницької спільноти. Перед публікацією ці дані були анонімізовані: імена користувачів AOL, IP-адреси та інші ідентифікатори були вилучені. Кожному користувачеві та його пошуковим запитам було присвоєно випадкове число. У наборі даних виявилося деяке унікальне поєднання запитів, що дозволило двом журналістам з New York Times швидко ідентифікувати одного з користувачів. Пошукові запити, такі як «ландшафти в Лілберні, штат Джорджія», «кілька людей з прізвищем Арнольд» та «будинки, продані в тіньовому озері підрозділу округу Гвіннетт, штат Грузія», дозволили знайти пані Тельму Арнольд, 62-річну вдову, яка проживала в Лілберні, штат Джорджія [11], яка пізніше в інтерв'ю підтвердила пошукові запити. Цей випадок привів до вилучення

опублікованого набору даних та змусив власників даних вжити заходів щодо обмеження доступу до таких інформаційних ресурсів у майбутньому.

Ці приклади демонструють, що проблема захисту конфіденційності даних є складною і потребує комплексного підходу до розробки стратегій анонімізації та захисту даних.

1.4 Постановка задач дослідження

Метою даної роботи є дослідження існуючих методів анонімізації даних і, на їх основі, розробка власного методу, який би задовільняв умовам анонімізації даних в медичних системах. Згідно з поставленою метою у роботі потрібно розв'язати наступні задачі:

В ході дослідження потрібно розробити:

– алгоритми, які забезпечать анонімізацію багатовимірних медичних даних. Ці алгоритми повинні ефективно замінювати ідентифікатори на неідентифікуючі дані, забезпечуючи при цьому збереження корисної інформації для подальшого використання;

– метрики якості анонімізації, які дозволять оцінити рівень захисту конфіденційності та відновлюваність корисної інформації. Це включає оцінку ступеня деградації даних та визначення можливості відновлення особистої інформації;

– програмне забезпечення або інші технічні засоби для впровадження розроблених алгоритмів анонімізації в реальних медичних системах.

Для досягнення поставлених задач необхідно провести детальний аналіз існуючих методів анонімізації, щоб виявити їх сильні та слабкі сторони. На основі отриманих даних буде розроблено новий метод, який оптимально поєднуватиме захист конфіденційності з функціональністю. Особлива увага приділятиметься забезпеченню високої швидкості та

ефективності алгоритмів анонімізації, що важливо для використання в медичних установах з великими обсягами даних. Крім того, важливо забезпечити сумісність нового методу з існуючими медичними системами та стандартами захисту даних. Очікується, що розв'язання цих задач дозволить забезпечити захист конфіденційності медичних даних та збереження їхньої корисної інформації для подальшого аналізу та використання в інтелектуальних медичних системах.

2 МЕТОДИ АНОНІМІЗАЦІЇ ДАНИХ

2.1 Анонімізація та конфіденційність даних

Анонімізація означає повне видалення ідентифікаторів особи, а також узагальнення будь-яких інших даних, які можуть бути використані для встановлення зв'язків з особою. Хоча такі нормативні акти, як GDPR, можуть вимагати анонімізації даних, чітких вказівок щодо цього не існує. Анонімізація відрізняється від деідентифікації, яка стосується видалення або заміни персональних ідентифікаторів у наборі даних таким чином, що зв'язок між особою та її записом даних може бути встановлений лише уповноваженою третя сторона.

Три основні цілі анонімізації є збереження:

– корисності даних (вимірюється кількістю втрат, спричинених технікою анонімізації, наприклад, втратою інформації);

– конфіденційності (вимірюється відповідністю даних обмеженням моделі конфіденційності);

– правдивість даних (кожен анонімізований запис відповідає єдиному запису в оригінальній таблиці).

Поняття конфіденційності часто пов'язують з реляційною моделлю даних в якій скінченні набори даних організовані в таблиці (відношення), які складаються зі стовпців (атрибутів) і рядків (записів). У контексті конфіденційності атрибути можна класифікувати на прямі ідентифікатори (DID), квазіідентифікатори (QID) та чутливі атрибути (SA). DID однозначно ідентифікують особу, наприклад, номер соціального страхування або водійське посвідчення. QID самі по собі не можуть ідентифікувати особу, але в поєднанні можуть повторно ідентифікувати особу. SA фізичної особи особи повинні бути приховані від потенційних супротивників, в той час як несуттєві атрибути можуть бути оприлюднені і вважаються такими, що вже вже відомі супротивникам. Оскільки

ідентифікатори особи є важливим компонентом різних моделей анонімізації, ми формально визначаємо QID наступним чином:

Заданий універсум U та таблиця $T(A_1, ..., A_m)$, що містить набір даних, що стосується певного набору осіб, нехай $f: U \rightarrow T$ та $g: T \rightarrow U' \subset U$. Тоді QID – це набір нечутливих атрибутів $Q_T = \{A'_j\}_{1 \leq j \leq r} \subset \{ A_1, ..., A_m\}$, для яких одну особу можна повторно ідентифікувати, коли вони комбінуються, тобто $\exists p \in U: g(f(p_i)[Q_T]) = p_i$.

Розрізнення прямих DID від QID полягає у тому, що DID прямо вказують на конкретних осіб або елементи в базі даних, тоді як QID зазвичай відносяться до атрибутів, що можуть бути використані для непрямого ідентифікування осіб (рисунок 2.1). Доцільне використання кожного з цих типів ідентифікаторів залежить від потреб аналізу даних та вимог до захисту приватності [4].

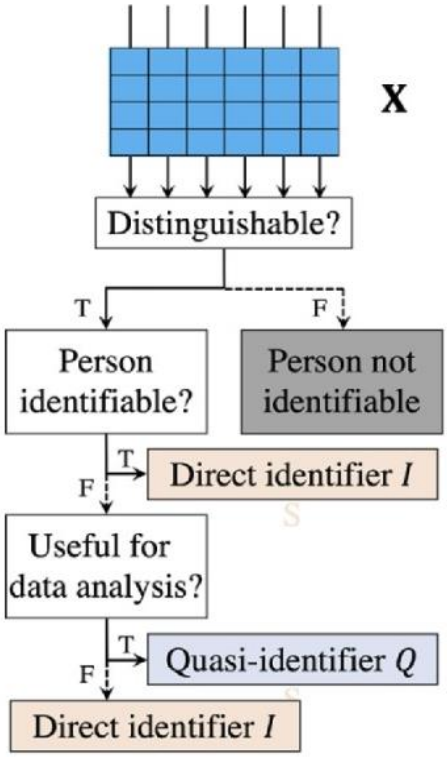

Рисунок 2.1 – Різниця між QID та DID

Існують три основні моделі анонімізації наведені на рисунку 2.2.

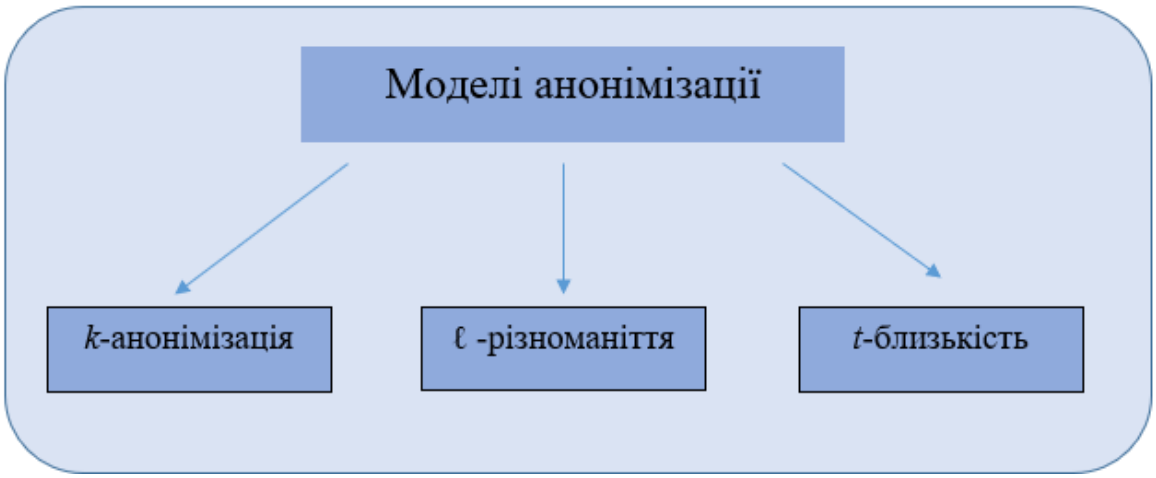

Рисунок 2.2 – Моделі анонімізації

k-анонімізація вимагає, щоб принаймні k осіб мали однакові атрибути. Оскільки QID містить поля, які, ймовірно, з'являться в інших відомих наборах даних, *k*-анонімізація гарантує, що кожна особа залишається анонімною в межах своєї відповідної групи (класу еквівалентності).

Нехай Q_T є QID для таблиці $T(A_1, ..., A_m)$, тоді T задовольняє *k*-анонімність тільки тоді, коли є принаймні k ідентичних записів для кожної унікальної комбінації $t[Q_T]$ значень записів щодо QID.

Наприклад, якщо k = 10, то кожен клас еквівалентності повинен мати принаймні 10 схожих записів. Це гарантує, що зловмисник не зможе визначити ідентичність одного запису. Однак, коли k занадто велике, корисність знижується. Крім того, відсутність достатньої гетерогенності чутливих атрибутів обмежує конфіденційність, яку пропонує модель *k*-анонімності.

ℓ-різноманіття долає обмеження *k*-анонімізація, враховуючи різноманітність серед SA. Різноманітність гарантує, що в кожному класі еквівалентності є щонайменше ℓ -різноманітні значень SA у кожному класі еквівалентності.

Блок q* є ℓ-різноманітним, якщо він містить принаймні ℓ різних значень для SAS, і ℓ найчастіших значень мають приблизно однакову частоту. Таблиця є ℓ-різноманітною, якщо кожний блок q* є ℓ-різноманітним.

Але ℓ -різноманіття не може запобігти атакам на розкриття атрибутів.

Наприклад, хвороба є SA і що таблиця є 3-різноманітною (тобто 3 різних значення SA в кожному класі еквівалентності). Для класу еквівалентності, де значеннями атрибутів виразка шлунка, гастрит і рак шлунка, противник може зробити висновок, що людина має проблеми, пов'язані зі шлунком, оскільки всі три захворювання в класі еквівалентності пов'язані зі шлунком.

t-близькість гарантує, що відстань між розподілом чутливих значень у кожному класі еквівалентності та вихідному класі не перевищує порогового значення t. Отже, чим менше значення t, тим вищий рівень конфіденційності.

Нехай P_S буде розподілом SA у таблиці T, а Q_S – розподілом цього атрибута у блоку q*. Блок q* вважається таким, що має *t*-близькість, якщо відстань d між P_S і Q_S не перевищує t для будь-яких SA. Таблиця T має *t*-близькість, якщо це виконується для всіх її блоків q*.

t-близькість долає обмеження *ℓ*-різноманіття за наявності викривленого розподілу ознак, в якому домінує одне чутливе значення. Забезпечення *t*-близькості для наведеного вище прикладу означатиме, що для таблиці, яка задовольняє 3-різноманітності, кожен клас еквівалентності буде містити не всі проблеми, пов'язані зі шлунком, але й інші типи захворювань, такі як пневмонія. Поширені функції відстані, що використовуються для вимірювання близькості, включають KL та EMD. На основі цих моделей формуються всі методи анонімізації, які мають на меті забезпечити ефективне та безпечне захист особистих даних у контексті ШІ та МН.

2.2 Методи анонімізації та псевдоанонімізації

Методи анонімізації та псевдоанонімізації (рисунок 2.3) набули значної популярності в епоху великих даних і посилення законодавчих вимог щодо захисту персональної інформації, таких як GDPR в Європейському Союзі. Вони дозволяють організаціям використовувати дані для аналізу та досліджень без порушення приватності користувачів. Основні техніки анонімізації включають в себе агрегацію даних, де індивідуальні дані об'єднуються в групи, та усереднення даних, де особисті характеристики замінюються середніми значеннями. Ще одним методом є випадкове перемішування даних, що порушує зв'язок між даними і реальними особами.

Псевдоанонімізація часто використовує криптографічні методи для заміни реальних ідентифікаторів на псевдоніми або коди, які можуть бути зворотно розшифровані за наявності спеціального ключа. Це дозволяє відновити вихідну інформацію у разі необхідності, наприклад, для медичних досліджень або правових розслідувань. Використання псевдоанонімізації також полегшує обмін даними між організаціями, зберігаючи при цьому захист конфіденційності.

Важливим аспектом при застосуванні цих методів є забезпечення достатнього рівня захисту від деанонімізації, тобто можливості відновлення ідентифікаторів з анонімізованих або псевдоанонімізованих даних.

Проте навіть найбільш досконалі методи анонімізації та псевдоанонімізації мають свої обмеження і ризики. Зокрема, комбінування анонімізованих даних з іншими доступними даними може призвести до повторної ідентифікації осіб. Тому важливо регулярно оцінювати ефективність застосовуваних методів і оновлювати їх відповідно до нових загроз та технологічних досягнень.

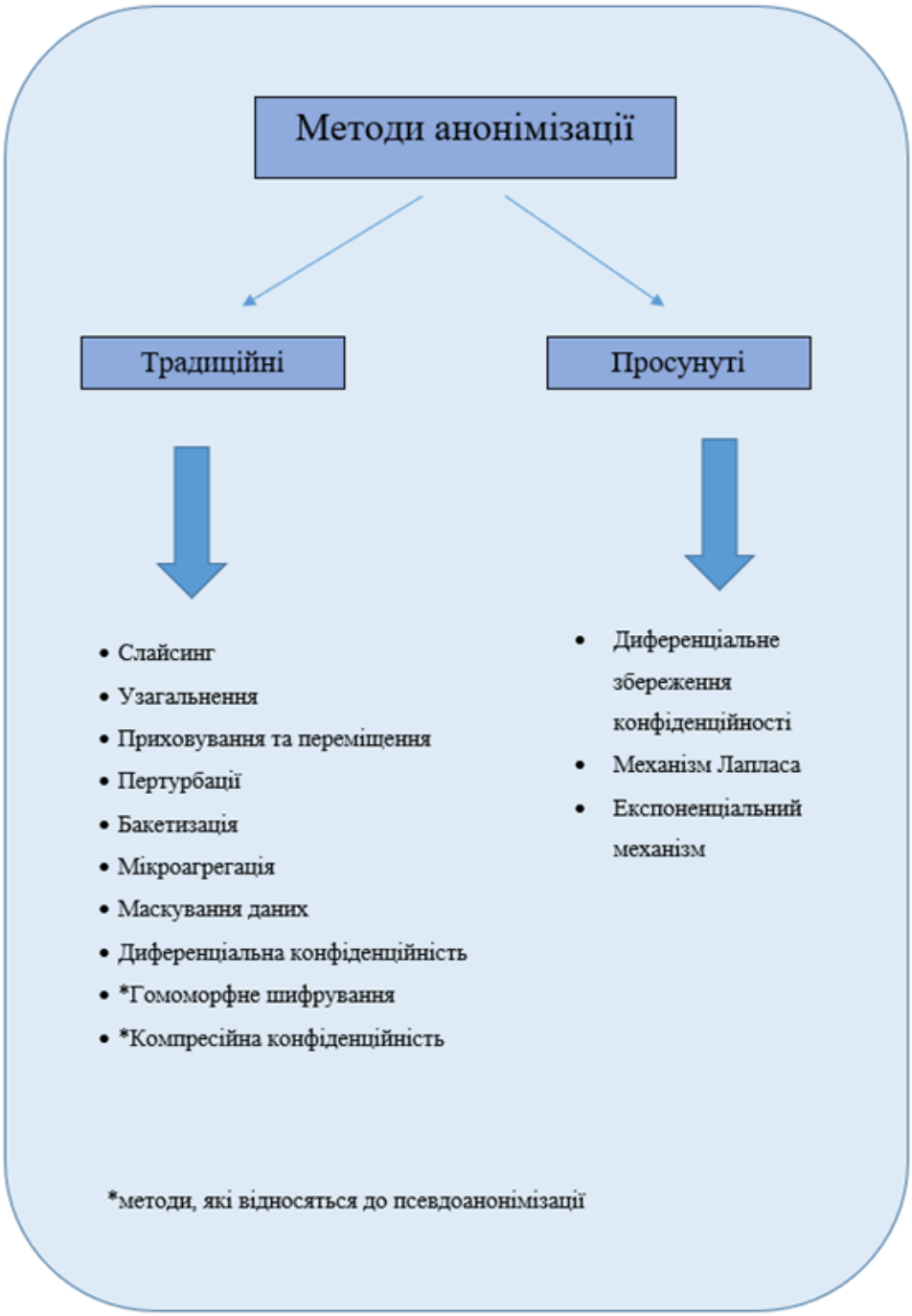

Рисунок 2.3 – Методи анонімізації

2.2.1 Традиційні методи анонімізації

Слайсинг: передбачає поділ даних на групи як по горизонталі, так і по вертикалі. Спочатку виконується вертикальний зріз за допомогою розбиття атрибутів на стовпчики, де кожен стовпчик містить підмножину атрибутів.

Потім виконується горизонтальний зріз шляхом розбиття кортежів на корзини, де кожна корзина містить підмножину кортежів. Метою зрізу є забезпечення того, щоб атрибути, які мають високу кореляцію, були згруповані разом.

Узагальнення: замінює QID значеннями іншими, менш специфічними значеннями, які узгоджуються з вихідними даними. Більшість алгоритмів узагальнення використовують глобальне перекодування або повне доменне узагальнення значень атрибутів. Це означає, що однакове перетворення застосовується до кожного значення QID. Однак, корисність даних може бути може бути вищою, якщо до кожного класу еквівалентності застосовується локальне перекодування. Анонімізація, досягнута за допомогою підходів, що ґрунтуються лише на узагальненні, неминуче спотворює записи. Тому надмірне узагальнення негативно впливає на аналіз анонімізованого набору даних.

Приховування та переміщення: для того, щоб зменшити надмірне узагальнення, використовуються методи приховування та переміщення. Приховування передбачає видалення пропусків, а переміщення передбачає зміну QID пропусків. Пропуски є основною причиною надмірного узагальнення оскільки пропуски віддалені від інших записів, і оскільки їх небагато, недостатньо згрупувати їх в однакові еквівалентні класи. У цих методах значення пропусків вилучаються або змінюються.

Приховування може виконуватись на рівні записів або на рівні клітинок. Приховування на рівні записів призводить до надмірного видалення записів для еквівалентних класів, які не відповідають обмеженням конфіденційності, тоді як придушення на рівні клітинок видаляє лише ті значення QID, які роблять кортеж таким, що порушує обмеження моделі конфіденційності.

Таким чином, це призводить до зменшення втрат інформації. Приховування стає в нагоді при використанні разом з моделями машинного навчання, тому що приховані значення можна обробляти як відсутні. Однак,

у великих наборах даних наборах даних повне витіснення стовпців може потенційно зашкодити корисності

Пертурбації: замість приховування, методи пертурбацій можуть використовуватись для посилення узагальнення. Пертубрація передбачає заміну чутливих значень та QID підробленими масками оригінальних даних. Використання пертурбацій для досягнення анонімізації призводить до кращої корисності, оскільки ступінь узагальнення обмежується шляхом вставки фальшивих записів до еквівалентних класів.

Бакетизація: відокремлює чутливі атрибути від QID шляхом випадкової перестановки або заміни місцями значень чутливих атрибутів у кожній корзині. Це досягає кращої корисності, ніж узагальнення. Однак, оскільки при бакетизації значення QID публікуються в їхніх оригінальних формах, зловмисник все одно може ідентифікувати особу за допомогою QID, а отже не перешкоджає розкриттю членства. Для кращої конфіденційності, узагальнення та бакетизація можуть бути застосовані до одного і того ж набору даних для спільного досягнення *k*-анонімності та ℓ-різноманіття відповідно.

Мікроагрегація: забезпечення того, що один і той самий набір даних задовольняє *k*-анонімність, ℓ-різноманіття та *t*-близькості призводить до значних втрат інформації через узагальнення, приховування та пертурбації навіть можуть бути недосяжним на практиці. Алгоритм мікроагрегування, який враховує числові QID для створення *k*-анонімних еквівалентних класів шляхом заміни чутливих атрибутів на замасковані значення. Мікроагрегація передбачає заміну значень груп з *k*-найближчих записів їхнім центроїдом. Оптимальним розв'язком цієї задачі є NP-hard.

Маскування даних: ця техніка часто використовується для створення структурно схожої, але неавтентичної версії необроблених даних. Ця техніка допомагає захистити оригінальні конфіденційні дані, забезпечуючи функціональну заміну, і її слід використовувати в налаштуваннях, у яких оригінальні необроблені дані не потрібні.

Диференціальна конфіденційність (DP): спрямована на те, щоб допомогти організаціям краще зрозуміти вимоги кінцевих користувачів шляхом максимізації точності пошукових запитів і мінімізації ймовірності ідентифікації інформації персональних даних. На практиці це досягається за допомогою таких методів, як фільтрація даних, адаптивна вибірка, додавання шуму шляхом фазифікації певних функцій, а також аналіз або блокування нав'язливих запитів. По суті, алгоритм DP оновлює значення, залишаючи одні недоторканими, а інші замінюючи таким чином, що потенційний зловмисник не зможе визначити, чи є значення підробленим чи справжнім.

Гомоморфне шифрування (HE): ця техніка дозволяє виконувати обчислення безпосередньо із зашифрованими даними, без необхідності розшифровувати дані. Недоліком такого методу є низька швидкість виконання. Методика, яка використовує метод шифрування, не може розглядатися як анонімізація. Наявність ключа робить дані теоретично оборотними і, отже, є псевдоанонімізацією даних. Добре відоме розширення HE називається адитивним HE, яке підтримує безпечне додавання чисел, враховуючи лише зашифровані дані.

Компресійна конфіденційність (CP): це техніка, яка пропонує виконати приватизацію шляхом відображення вихідних даних у просторі з меншою розмірністю. Зазвичай це досягається шляхом вилучення ключових функцій, необхідних для моделі машинного навчання, перед надсиланням даних на хмарний сервер. З цією метою власники даних контролюють конфіденційність. Крім того, цю техніку можна виконати перед застосуванням обраних моделей конфіденційності. По суті, її можна розглядати як техніку зменшення розмірності, яка також зберігає конфіденційність [12].

Перевагою використання методів анонімізації перед методами псевдоанонімізації є те, що анонімні дані неможливо ідентифікувати, тоді як псевдоанонімні дані можна ідентифікувати.

Мета задоволення різних рівнів анонімності зазвичай досягається за допомогою двох послідовних кроків: вимірювання та оцінювання. Перше відноситься до використання звичайних моделей конфіденційності або моделей конфіденційності на основі машинного навчання для виконання анонімізації даних, а останнє є процесом оцінки ризику повторної ідентифікації та ступеня зручності використання анонімних даних. Операції анонімізації зазвичай застосовуються звичайними моделями конфіденційності або моделями на основі машинного навчання.

2.2.2 Просунуті методи анонімізації

Диференціальне збереження конфіденційності: методи диференціальної конфіденційності не збільшують і не зменшують вихідні дані про зміни індивідуальної інформації в базі даних. Припустимо, що Y – випадковий алгоритм, а Q – можливий результат для двох наборів P і R:

$$\Pr[Y[P]\varepsilon S \leq \exp(\varepsilon) \times \Pr[Y[R]\varepsilon S]. \quad (2.1)$$

Тоді алгоритм Y дає (ε) диференціальну конфіденційність ($\varepsilon > 0$). Механізм Лапласа: Лапласа та експоненціал є основною концепцією диференціального збереження конфіденційності. Крім того, рандомізація перевіряється в обох концепціях за допомогою аналізу чутливості. Для запиту Q чутливість виражається як

$$\Delta Q = \max_{P,R} \|Q(P) - Q(R)\|. \quad (2.2)$$

Він також використовує перетворення Лапласа для числового виведення, додаючи незалежний шум. Для Q: набір даних P у діапазоні W

метод M забезпечує ε-диференціальну конфіденційність за допомогою наступного рівняння:

$$M(P) = Q(P) + Lap\left(\frac{\Delta Q}{\varepsilon}\right). \quad (2.3)$$

Експоненціальний механізм: забезпечує рандомізовані результати для нечислових запитів, об'єднаних із функцією оцінки, щоб перевірити якість результату S. Припустимо, що c (P, θ) – функція оцінки набору даних P , яка оцінює вихідну якість $\theta \in \lambda$, а ΔQ – чутливість θ . Експоненціальний механізм задовольняє ε – диференціальну конфіденційність за наступним рівнянням [5]:

$$M(P) = \left(return\ \theta\ \alpha\ exp\left(\frac{eq(P,\theta}{2\Delta Q}\right)\right). \quad (2.4)$$

Захист і безпека даних є найважливішими аспектами платформи великих даних, і якщо дані не захищені заходами безпеки, дані можуть бути легко скомпрометовані. Щоб мати змогу працювати з величезною кількістю даних, розроблені методи глибокого навчання (DL).

Метод захисту даних пропонує певні заходи контролю над даними, що відображаються. Крім того, переваги підходів DL не лише обмежуються аналізом даних, але також знають класифікацію зображень, аналіз мови та тексту.

Таким чином, потреба в диференційованому захисті даних у DL зростає експоненціально, завдяки чому захист даних гарантується без втрати інформації. Цей підхід забезпечує надійний захист даних, що є важливим для забезпечення приватності та конфіденційності великих обсягів інформації.

2.3 Атаки на дані

Хакерські атаки на медичні дані (рисунок 2.4) можуть викликати витік конфіденційної інформації або призвести до блокування роботи систем, створюючи загрозу для пацієнтів. Викрадення медичних даних може слугувати для здійснення шахрайства, ідентифікаційної крадіжки та інших незаконних цілей, порушуючи приватність і безпеку пацієнтів, детільніше можна прочитати в опублікованій у Всеукраїнській студентській науково-практичній конференції «Гібридні війни сучасності: стійкість та протидія гібридним загрозам» статті, у додатку А на риснку А.1 наведено сертифікат.

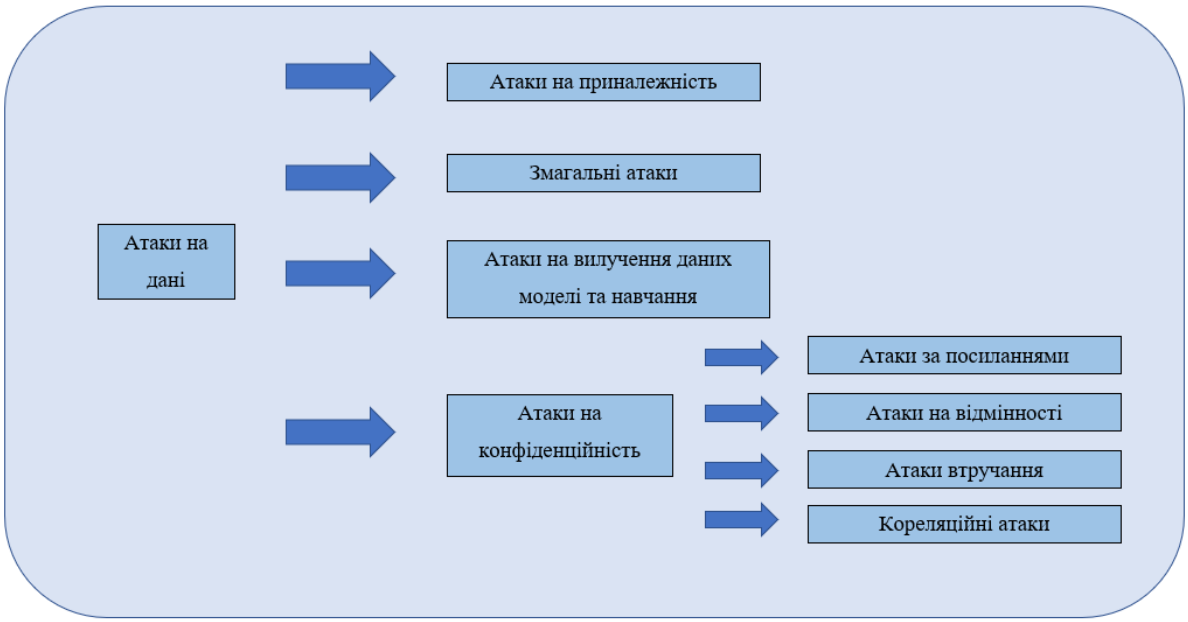

Рисунок 2.4 – Атаки на дані

Атаки на приналежність: викликані різною поведінкою навчальних і тестових наборів даних. Опонент робить висновок, чи присутній набір даних у наборі навчальних даних. Під час такої атаки зловмисник запитує ціль із набором даних, щоб отримати прогноз, представлений у вигляді вектора. Вектор прогнозування на тесті та навчальний набір даних представлений у них і використовується для навчання моделей атак. Ця

модель атаки надає ймовірності приналежності цільового набору даних до навчального набору даних (рисунок 2.5).

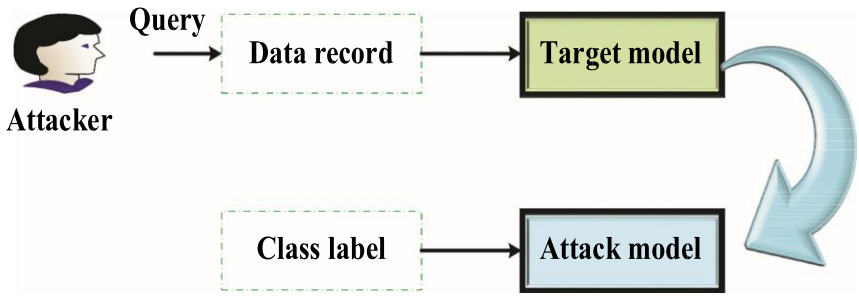

Рисунок 2.5 – Візуалізація атаки на приналежність

Змагальні атаки: змінюють оригінальне зображення та навмисно заважають глибокій нейронній мережі шляхом плутанини. У змагальній атаці заважаючий шум додається до вихідного вхідного зображення, щоб створити змагальний результат (рисунок 2.6). Ця атака впливає на точність класифікації глибоких нейронних мереж.

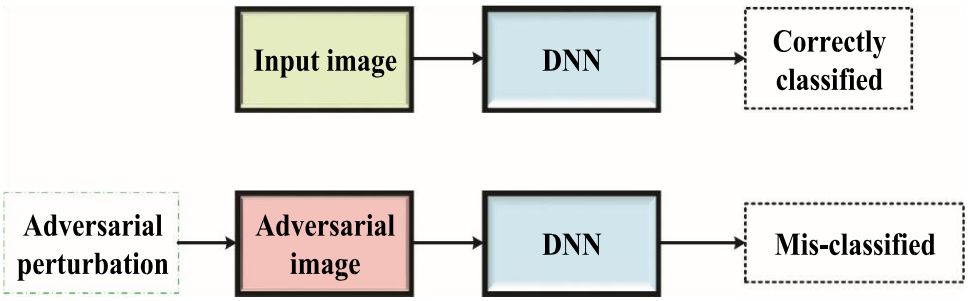

Рисунок 2.6 – Візуалізація змагальної атаки

Атаки на вилучення моделі: витягують параметри, навчені в моделі, з персональних даних, у яких зловмисник дублює функціональність моделі. Через витік параметрів розкривається конфіденційність навчального набору.

Під час передачі даних у системі відбувається кілька атак на конфіденційність [13], інформацію про які було опубліковано в Міжнародному молодіжному форумі «Радіоелектроніка і молодь у XXI столітті», у додатку А на рисунку А.2 наведено диплом.

При публікації інформаційних таблиць про певну групу осіб, важливо дбати про збереження конфіденційності цих осіб, чиї дані надаються. У цьому контексті конфіденційність означає два аспекти: перше, жодна особа, що згадується в опублікованих даних, не повинна бути ідентифікована; друге, будь-яка інформація, яка вважається конфіденційною, не повинна бути пов'язана з конкретною особою.

Існують два типи загроз, які можуть підіймати питання про конфіденційність осіб, чиї дані публікуються. Якщо процес анонімізації не враховує можливості зібрати інформацію з інших джерел про осіб, які згадуються в опублікованих даних, це може створювати умови для атак на узгодження записів та атрибутів. Такі атаки, зазвичай, спрямовані на таблиці реляційних даних, де кожний рядок відображає унікальну особу, а кожний стовпець містить атрибут, що описує певний тип інформації про цих осіб у таблиці. Додатковим ризиком є можливість комбінування анонімізованих даних з іншими відкритими джерелами, що може допомогти в ідентифікації окремих осіб або розкритті конфіденційної інформації.

Узгодження записів: атака, має на меті однозначну ідентифікацію групи осіб у опублікованій таблиці даних за допомогою раніше зібраної інформації про цих осіб. Суїні [7] описав сценарій атаки, при якій зловмисник використовує заздалегідь зібрану інформацію про групу осіб з опублікованої таблиці даних, отриману зовнішнім джерелом. Ці дані, відомі як фонові знання зловмисника, використовуються для ідентифікації конкретної особи в опублікованій таблиці. Поєднання певних ознак, що відомі як QID, у таблиці може призвести до ідентифікації групи осіб. Атрибути QID є загальними між опублікованими та зовнішніми даними. Це

дозволяє зловмиснику використовувати ці атрибути для зіставлення та ідентифікації окремих осіб у наведеній таблиці.

Приклад 1. Нехай лікарня передає дані своїх пацієнтів дослідницькому центру для аналізу. У таблиці 2.1 представлена опублікована, але не оброблена, інформація після видалення явних ідентифікаторів. Атрибути «Рік народження», «Стать» та «Професія» є QID в цьому випадку. Як приклад, врахуймо, що зловмисник виявив з анонімної зовнішньої таблиці, що його сусід Дмитро має значення {2000, Ч, Біолог} для цих атрибутів QID. Така комбінація ідентифікує пацієнта в записі 1. Таким чином, атакуючий може встановити, що запис 1 у таблиці 2.1 належить Дмитру і що він має діагноз ВІЛ.

Таблиця 2.1 – Візуалізація прикладу 1, необроблена таблиця

№	Рік народження	Стать	Професія	Діагноз
1	1990	Ч	Біолог	ВІЛ
2	1991	Ч	Лікар	ОРВІ
3	1990	Ч	Програміст	ВІЛ
4	2000	Ч	Архітектор	ОРВІ
5	1997	Ж	Архітектор	ОРВІ
6	1998	Ж	Програміст	ВІЛ
7	1998	Ж	Програміст	ВІЛ

Узгодження атрибутів: атака, має на меті пов’язування конфіденційних фрагметнів з певними особами.

Залежно від серйозності описаної інформації, атрибут вважається чутливим постачальником даних. Наприклад, зарплата, хвороба та РНОКПП – це приклади чутливих ознак. Зауважимо, що не всі значення домену чутливого атрибута є чутливими. Деякі значення можуть бути відзначені як нечутливі. Чутливі значення не повинні бути пов'язані з конкретними особами. Наприклад, у таблиці 2.1 атрибут «Діагноз» = {ВІЛ, ОРВІ} вважається чутливим.

Проте у деяких випадках «ОРВІ» може вважатися нечутливим. Навіть якщо запис пацієнта не є унікальним у таблиці, зловмисник може визначити хворобу цільової жертви. Ми оцінюємо ступінь визначеності зловмисника, виводячи чутливе значення жертви.

Приклад 2. Розглянемо записи пацієнтів у таблиці 2.1. Припустимо, що зловмисник знає, що запис про цільового пацієнта Марина містить {1998, Ж, Програміст}. Навіть якщо існують два записи з однаковими значеннями атрибутів QID, обидва мають «ВІЛ» на чутливому атрибуті «Діагноз». Тому зловмисник може з 100% впевненістю вважати, що Марина має ВІЛ, що означає, що довіра до висновку зловмисника становить 100%.

Атака мінімальності. Якщо не приділяється достатньо уваги захисту конфіденційності людей, їх опублікована таблиця даних залишається вразливою до атак на узгодження записів та атрибутів. В результаті, оригінальна таблиця даних повинна бути перетворена в анонімну версію, яка відповідає деяким вимогам конфіденційності і, таким чином, запобігає цим атакам.

Вонг та ін. [14] показали, що навіть після анонімізації таблиці даних зловмисник все ще може пов'язувати чутливі значення з відповідними особами. Можливість такої атаки показано у наступному прикладі.

Приклад 3. Розглянемо просту версію даних про пацієнтів у таблиці 2.2, де «Стать» та «Професія» є атрибутами QID.

Таблиця 2.2 – Візуалізація прикладу 3, необроблена таблиця

№	Стать	Професія	Діагноз
1	Ч	Програміст	ОРВІ
2	Ч	Програміст	ОРВІ
3	Ч	Програміст	ОРВІ
4	Ж	Архітектор	ВІЛ
5	Ж	Архітектор	ВІЛ

Нехай «ВІЛ» є єдиною чутливою цінністю для хвороби. Під атакою на узгодження з атрибутами таблиця 2.2 у своїй оригінальній версії з 100% впевненістю виявляє, що якщо цільова особа є жінкою, архітектором, то вона має ВІЛ. Щоб знизити достовірність цього висновку, можна використовувати методи анонімізації, які змінюють або приховують певні атрибути, такі як стать чи професія, зберігаючи при цьому корисну інформацію для аналізу даних.

Таблиця 2.3 містить найбільші загальні значення для всіх атрибутів QID, і це призводить до наявності лише однієї групи QID {М-Ж, Працюючий}, в якій зловмисник може зробити з максимум 40% впевненості висновок про те, що пацієнт має ВІЛ.

Таблиця 2.3 – Візуалізація прикладу 3, анонімізована таблиця

№	Стать	Професія	Діагноз
1	Ч-Ж	Працюючий	ОРВІ
2	Ч-Ж	Працюючий	ОРВІ
3	Ч-Ж	Працюючий	ОРВІ
4	Ч-Ж	Працюючий	ВІЛ
5	Ч-Ж	Працюючий	ВІЛ

Отже, таблиця 2.2 анонімізована до таблиці 2.3 відповідно до набору дерев таксономії на рисунку 2.7.

Оскільки таблиця 2.3 містить загальну інформацію та має обмежену достовірність логічних висновків, ефективно захищена від атак на узгодження атрибутів. І хоча Вонг та інші [14] зазначали, що це може не завжди бути правдою, їхні доводи розширюють знання зловмисника. Це стосується не лише зовнішніх даних, аналогічних таблиці 2.2, за винятком атрибута «Діагноз», але і алгоритму анонімізації та набору дерев таксономії.

Отже, зловмисник отримує такі відомості:

– зовнішня таблиця даних, схожа на таблицю 2.2, але без атрибута «Діагноз»;

– вимога обмеження довіри до висновку про ВІЛ до 70%, як загальновідомої інформації;

– опублікована анонімна таблиця, а саме таблиця 2.3;

– набір дерев таксономії, які проілюстровані на рисунку 2.7.

З цими відомостями зловмисник визначає, що таблиця 2.3 була анонімізована. Таким чином, зловмисник може визначити, чи порушуються вимоги конфіденційності групою, що включає записи з 1 по 3, чи групою з 4 по 5. Наприклад, якщо перша група містить двох пацієнтів з ВІЛ, то довіра до неї складає 2/3 = 66,67%, що відповідає вимогам конфіденційності, і анонімізація не є необхідною. Однак, якщо друга група (записи 4-5) також містить двох пацієнтів з ВІЛ, то довіра до неї становить 100%. Зловмисник робить висновок, що анонімізація була проведена, оскільки в групі з записів 4-5 мають бути пацієнти з ВІЛ. В іншому випадку анонімізація не є необхідною.

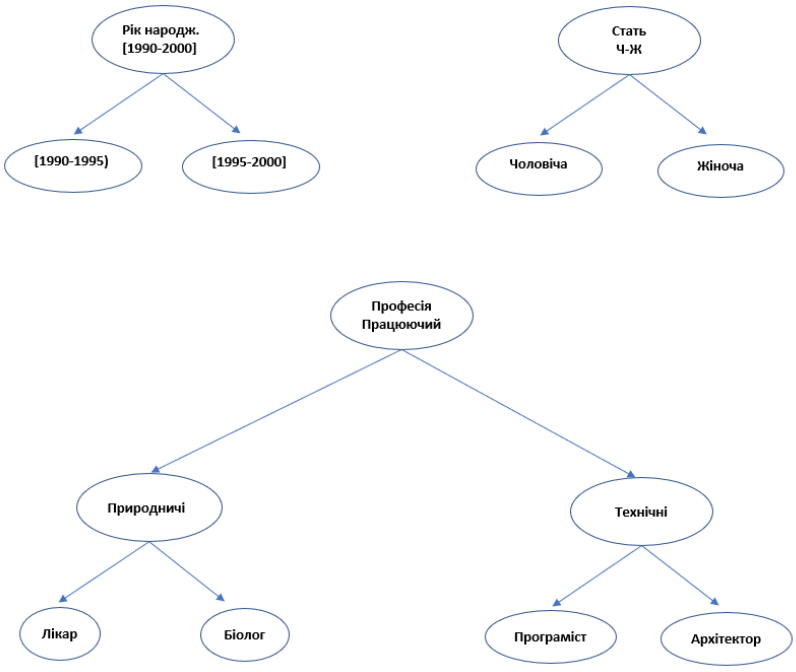

Рисунок 2.7 – Дерева таксономії для даних пацієнтів

Описаний вище приклад ілюструє типовий сценарій атаки мінімальності [14], який використовує той факт, що існуючі алгоритми анонімізації дотримуються принципу мінімальності при перетворенні таблиці даних [15]. Суть цього принципу полягає в тому, що необроблена таблиця даних перетворюється на її найменш анонімну версію, оскільки анонімізація вже втрачає деяку інформацію, і немає необхідності у подальшій анонімізації на вже анонімній таблиці, яка вже відповідає вимогам конфіденційності.

Атака за посиланнями: щоб зберегти конфіденційність даних під час спілкування, власник повинен анонімізувати дані, щоб приховати конфіденційну інформацію. В еру великих даних людей ідентифікують за кількома ознаками. З іншого боку, зловмисник пов’язує анонімні дані з неанонімними даними в різних наборах даних, спричиняючи порушення конфіденційності. Пов'язування зовнішніх даних з анонімними даними називається атакою за посиланнями.

Атака на відмінності: механізм обмеження запитів уникає запитів до окремої особи, таким чином агрегуючи запити, а не надсилаючи запит до окремої особи. У той же час, довільні запити в різні моменти можуть викликати розбіжності

Атака втручання: передача індивідуальних даних має бути дезінфікована, щоб запобігти доступу зловмисників до конфіденційної інформації. Зловмисник може застосувати методи інтелектуального аналізу даних, щоб отримати особисту інформацію незаконним шляхом. Цей тип атаки називається втручанням, він може спричинити витік даних, якщо зловмисник здогадається про їх реальну цінність.

Кореляційна атака: дані реального світу мають високу кореляцію між даними, розкриваючи більше інформації, ніж очікувалося. Зловмисник із кореляційною інформацією може мати більший обсяг інформації про конфіденційність і, отже, може мати вищі шанси на порушення даних.

Базові знання супротивника можуть генерувати кореляційну інформацію. Потім зловмисник створює кореляційну атаку, щоб порушити конфіденційність даних.

Щоб уникнути цих атак в епоху великих даних, використовуються підходи на основі *k*-анонімності, криптографії та алгоритмів оптимізації.

Алгоритми оптимізації ітераційно виконуються шляхом порівняння різних рішень, отриманих на кожній ітерації, доки не буде отримано оптимальне рішення для формування задачі. Таким чином, підходи, засновані на оптимізації, широко застосовуються в різних додатках. У зв'язку з цим алгоритми анонімізації застосовуються разом з алгоритмами оптимізації, щоб мінімізувати обчислювальну складність.

Алгоритми на основі анонімізації використовуються для захисту індивідуальних даних, зберігаючи при цьому зручність використання даних. Запропонований алгоритм Кеном Ейпогу та ін. був ініціалізований частотним аналізом квазіідентифікатора та калібруванням ключових значень. Після цього хаотична функція генерувала нові значення для вибраних ключових значень. Результати, отримані за допомогою цього алгоритму, демонструють, що запропонований алгоритм забезпечує конфіденційність особи перед публікацією даних.

Такі підходи є надзвичайно важливими в контексті зростання обсягу даних, що зберігаються та обробляються в сучасному світі. Крім того, вони сприяють створенню безпечних систем зберігання та обробки даних, що відповідають вимогам законодавства про захист персональних даних. Застосування цих методів також знижує ризики витоку конфіденційної інформації та підвищує довіру користувачів до технологічних рішень. В результаті, компанії та організації можуть більш ефективно використовувати зібрані дані для аналізу та прийняття рішень, не порушуючи при цьому приватність користувачів. Це підвищує загальну безпеку та стабільність інформаційних систем.

3 АНОНІМІЗАЦІЯ МЕДИЧНИХ ДАНИХ. ОГЛЯД РІШЕНЬ

3.1 Реляційний модель представлення даних

Традиційний спосіб представлення даних про стан здоров'я – реляційна модель даних. Щоб зменшити втрати інформації через узагальнення, проблему *k*-анонімності можна розглядати як задачу кластеризації або розбиття, де метою є знаходження наборів кластерів (тобто класів еквівалентності), кожен з яких містить принаймні *k* записів. Записи у кожному розділі або кластері потім узагальнюються таким чином, що всі вони мають однакове значення QID і містили принаймні k записів. Однак, коли є багато QID, техніка кластеризації техніка кластеризації все ще страждає від великих втрат інформації і не захищає від ℓ -різноманіття.

Маджид та ін. зосередилися на адаптивному узагальненні атрибутів на основі різного інформаційного вмісту QID [15]. Вони запропонували рішення проблеми «різноманітності та *t*-близькості в незбалансованому наборі даних» шляхом створення гнучкої ієрархії узагальнення на основі вразливості QID, щоб виявити ідентичність і різноманітність SA у кожному класі еквівалентності. Запропонований ними метод підвищує корисність шляхом контролюючи надмірне узагальнення QID, які є менш вразливими для виявлення ідентичності в різних класах.

Для того, щоб один і той самий набір даних був *k*-анонімним, ℓ-різноманітним і *t*-закритим, Гал. та ін. запропонували алгоритм мікроагрегації, який використовує числові QID для створення k-анонімних кластерів і замінює чутливі атрибути на замаскованими значеннями [16]. Інші методи, що використовуються для спільного досягнення всієї *k*-анонімності, ℓ-різноманітності та *t*-близькості для медичних даних, включають SHARE та CAHD.

Методи інкрементної анонімізації базуються на накопиченні, агрегуванні та кластеризації. Для цих методів кожен новий фрагмент даних є кортежем, що складається з QID та чутливих атрибутів. Коли надходять нові дані, вони проходять через механізм агрегації, який розподіляє дані по кластерах на основі втрати інформації. Ці кластери потім передаються до моделей конфіденційності. Якщо кортеж не задовольняє моделі конфіденційності, наприклад, *l*-різноманітності, до кортежу застосовується метод пертурбацій шляхом випадкового додавання деяких згенерованих атрибутів до вихідних даних.

Методи анонімізації з урахуванням корисності (UAA) використовують анонімізацію на рівні атрибутів, яка зберігає початкові значення QID на основі присвоєного значення корисності і застосовує локальне перекодування для анонімізації даних. Крім того, Пауліс та ін. запропонували метод *(k, km)*-анонімності, який має обмежену корисність і при цьому забезпечує конфіденційність як демографічної інформації, так і кодів діагнозів. Окрім використання популярних метрик, таких як втрата інформації, точність та розрізнюваність для вимірювання корисності, кілька робіт використовували точність навченої класифікаційної моделі як метрику корисності використовували точність навченої класифікаційної моделі. Такі підходи схиляються до використання машинного навчання до використання моделей машинного навчання або дерев рішень для вимірювання компромісу між точністю, анонімністю та конфіденційністю.

Оскільки дані про здоров'я можуть бути багатовимірними з декількома атрибутами, знання зловмисника про ці атрибути можна послабити і обмежити до L відповідно до LKC-моделі конфіденційності, запропонованої Мохаммедом та ін. Це означає, що не всі атрибути повинні бути анонімізовані, а лише комбінація L атрибутів, щоб задовольнити модель конфіденційності (наприклад, *k*-анонімність).

Анонімізація проти фонових атак. Маджид запропонував схему анонімізації, яка запобігає розкриттю особистих даних навіть у випадку,

коли супротивник має сильні фонові знання. Запропонований ними метод базується на перетворенні даних у фіксовані інтервали а потім заміни початкових значень середніми значеннями [16].

Атака на фонові знання: коли противнику відома деяка інформацію або QID про цільову особу, вона може реконструювати ідентифікаційну інформацію про особу. Така реконструктивна атака ставить під загрозу конфіденційність особи, яку атакують.

Атака на зв'язки: є однією з класичних атак на реляційні дані, коли зловмисник може повторно ідентифікувати або зв'язати запис в анонімному наборі даних, комбінуючи ідентифікатори QID з різних джерел з конкретною особою. Це вимагає певної форми атаки на фонові знання.

Атака на розкриття атрибутів: в атаці на розкриття атрибутів зловмисник має на меті отримати нову інформацію про SA. Зловмисник також може використовувати властивості QID, щоб оцінити SA. Зазвичай, розкриття атрибутів є побічним продуктом розкриття ідентифікаційних даних. Тим не менш, воно може бути агностичним до атаки на зв'язки. Наприклад, противник може бути зацікавлений у тому, щоб вивести SA, яка є спільною для всіх цільових осіб.

Атаки на розкриття членства: передбачають, що зловмисник намагається зробити висновок про наявність або відсутність особи в наборі даних. Наприклад, знаючи, що особа присутня в наборі даних про рак, можна зробити висновок, що вона хвора на рак, навіть якщо конкретний тип раку не може бути не можна визначити. Це може слугувати важелем для запуску атаки на розкриття атрибутів.

3.2 Графічні дані про стан здоров'я

Останнім часом електронні медичні картки представляють у вигляді графових даних для вилучення графів знань. Ці дані про здоров'я на основі графів надають багаті знання про взаємодію та основні властивості даних.

Наприклад, в мережі передачі захворювань або епідеміологічному графі, вузлами є пацієнти, тоді як ребра відображають взаємодію між пацієнтами. Тому і вузли, і ребра потрібно захищати, оскільки вони містять конфіденційну інформацію. Підхід до анонімізації даних про стан здоров'я на основі графів даних на основі графів можна досягти або видаленням ребер і міток вершин (наївна анонімізація) або додаванням нових ребер чи вузлів для зміни структури графа (структурна анонімізація).

Графовий варіант техніки *k*-анонімності, що використовується в реляційних даних про здоров'я, – це анонімність *k*-го ступеня.
Граф є *k*-ступенево анонімним, якщо для кожного вузла u у графі існує принаймні (*k* -1) інших вузлів з такою самою ступенем, як у вузла u.

Чжоу та Пей показали, що навіть коли конфіденційність особи зберігається за допомогою звичайних методів анонімізації, особа в графі може бути повторно ідентифікована через атаки сусідів. Це можливо в таких випадках, коли зловмисник має певні знання про сусідів жертви та зв’язки між ними. Щоб забезпечити конфіденційність, кожен вузол повинен мати k інших вузлів з таким же сусідством на відстані 1-го хопу.

Використовуючи техніку модифікації графів таку як додавання або видалення ребер, Лю та Терзі досягли *k*-ступеневої анонімності побудувавши послідовність k-анонімної ступеневої послідовності, де кожна вершина має такий самий ступінь, як і (*k*, ℓ) інших вершин [16]. Анонімна степенева послідовність потім використовується для побудови анонімізованого графа.

Мортазаві та Ерфані запропонували метод (*k*, ℓ)-анонімізації, який зберігає конфіденційність особи або вершини у графі таким чином, що навіть коли зловмисник знає максимум ℓ-сусідів вершини, він не може ідентифікувати цю вершину у групі з менш ніж k вершин [16]. Це досягається додавання ребер до графа, щоб задовольнити модель конфіденційності, а потім надлишкові ребра видаляються, щоб мінімізувати

зміни між анонімізованим і оригінальним графом. Однак, зміни можуть бути видимими для зловмисника.

Подібно до реляційних даних, підходи до анонімізації на основі кластеризації для графічних даних спочатку розбивають дані на кластери і будують графіки на основі даних у кожному кластері. Потім виконується модифікація графа, наприклад, видалення або додавання ребер, щоб задовільнити обмеження моделі конфіденційності. Желєва та Гетур показали, що чутливі зв'язки можуть бути виведені з анонімізованих графів. В якості методу захисту вони видалили всі ребра з анонімізованого графа і згорнули вершини в один вузол для кожного кластеру. Потім вони випадковим чином вибрали ребра з видалених ребер і призначили їх ребрами кластерів. Для того, щоб анонімізувати графи, зберігаючи при цьому структурну подібність, Фоффано та ін. застосували лему регулярності Семереді для розбиття кожної вершини на кластери та забезпечення, що ребра всередині розбиття поводяться майже випадковим чином. Таким чином, ребра між розбиттями залишаються незмінними, що робить його семантично подібним до оригінального графа [16].

Атака розкриття атрибутів вузла: атака схожа на атаки на зв'язки в реляційних даних, де QID може бути використаний для повторної ідентифікації особу. Вузол може бути пов'язаний з особою на основі набору атрибутів, призначених вузлу.

Атака по сусідству: оскільки вузли в графі пов'язані між собою, то зловмисник, маючи попередні знання про деяких сусідів може розкрити ідентичність інших пов'язаних вузлів. Наприклад, перед тим, як анонімізований граф буде опублікований, зловмисник може випадковим чином створити підграф і прикріпити його до певного цільового користувача. Атака вважається успішною, якщо зловмисник може знайти підграф в опублікованому анонімному графі. Він може потім виявити інші вершини, слідуючи по ребрах підграфа.

Атака на основі структури: це парасольковий тип атаки, що складається з декількох атак, де противник використовує унікальні структурні характеристики графа. До таких атак відносяться степенева атака (використання послідовності степенів графа для унікальної ідентифікації особи), атака на підграф (зловмисник знає підграф навколо цільового вузла), атака за відбитками пальців (зловмисник, знаючи відстань між концентратором і цільовим вузлом, може проникнути в цільовий вузол і цільовим вузлом, може втрутитися в приватність цільового вузла), а також атака на основі прогулянок (пошук на коротких дистанціях).

Атака на існування вузла: подібно до атаки розкриття членства в реляційних даних, зловмисник намагається виявити наявність або відсутність вузла в графі. Цій атаці зазвичай передує атака на основі структури.

3.3 Генерування синтетичних даних

Анонімізація даних через створення синтетичних даних – це процес заміни реальних даних у наборі даних іншими даними, які зберігають загальні статистичні властивості оригінальних даних, але не дозволяють ідентифікувати конкретних осіб або об'єктів. Цей підхід особливо корисний для захисту конфіденційності даних в тих випадках, коли оригінальні дані містять особисту або конфіденційну інформацію, і коли необхідно зберігати цю інформацію у вигляді даних для аналізу чи досліджень.

Основна ідея полягає в тому, що синтетичні дані створюються на основі оригінальних даних, але з додаванням певного рівня шуму або за допомогою інших технік генерації, які забезпечують анонімність даних. Це може включати створення синтетичних «схожих» даних, що відтворюють статистичні характеристики оригінальних даних, такі як розподіли, кореляції та інші важливі параметри.

Процес створення синтетичних даних може використовувати різноманітні методи, включаючи статистичні моделі, машинне навчання, техніки генеративних моделей (таких як генеративні згорткові мережі або варіаційні автокодери) та інші підходи.

Техніки деідентифікації, такі як *k*-анонімність, ℓ-різноманіття та *t*-близькість, не є стійкими проти атак повторного ідентифікування. Завдяки юридичним, конфіденційним та безпековим питанням, пов'язаним з медичними даними, і обмеженому доступу до них, сектор охорони здоров'я відстає від інших секторів у плані використання інформаційних технологій, обміну даними та міжсистемної взаємодії. Альтернативним методом є створення реалістичних синтетичних даних. Переваги використання синтетичних даних полягають у тому, що вони штучно створені, тому між реальними та синтетичними даними немає явного відображення. З цієї причини, на відміну від деідентифікованих даних, синтетичні дані залишаються стійкими до повторного ідентифікування. Якщо синтетичні дані можуть мати атрибути, подібні до фактичних даних, вони повинні допомогти компаніям та дослідникам у публічному використанні інформації без клопоту щодо отримання реальних даних.

3.4 Генеративні змагальні мережі

Генеративні змагальні мережі (GAN) – це модель машинного навчання, в якій дві нейронні мережі конкурують одна з одною, використовуючи методи глибокого навчання, для покращення своїх прогнозів. Зазвичай GAN працюють без нагляду і використовують для навчання кооперативну гру з нульовою сумою, де виграш одного гравця дорівнює втраті іншого [18]. GAN складаються з двох нейронних мереж: мережі-генератора (G) та мережі-дискримінатора (D), які навчаються конкурувати у генерації нових даних з тією ж статистикою, що й навчальний набір, наприклад, створюючи нові фотографії, які виглядають принаймні

достовірно для спостерігача, маючи багато реалістичних характеристик. На рисунку 3.1 показано приклад архітектури цієї нейронної мережі.

Завдяки двом конкуруючим мережам G та D GAN може генерувати синтетичні зображення, які майже неможливо відрізнити від реальних зображень. Використовуючи цю здатність створювати реалістичні синтетичні зображення, GAN були успішно застосовані в багатьох застосуваннях, таких як генерація зображень, синтез тексту-зображення, переклад зображення-зображення, генерація відео, музична генерація тощо. Усі ці роботи підтверджують, що GAN – це найкращий вибір для створення реалістичних синтетичних вибірок.

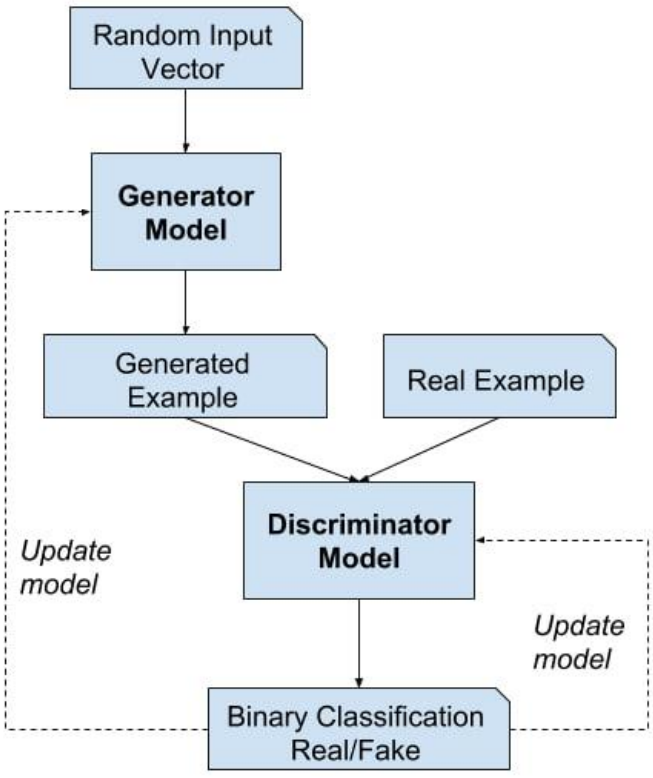

Рисунок 3.1 – Приклад архітектури моделі генеративної змагальної мережі

Впровадження GAN в різні сфери дозволяє ефективно вирішувати завдання, пов'язані з генерацією контенту, який має високу ступінь реалізму та корисності. Такий широкий спектр застосувань GAN свідчить про їхню

велику потенційну користь у багатьох областях, від творчого мистецтва до наукових досліджень. Використання GAN в різних сферах індустрії підтверджує їхню універсальність та ефективність у генерації реалістичних синтетичних даних.

3.5 Трансформери

Трансформери – це тип архітектури нейронних мереж, який став популярним у сферах обробки природної мови (NLP), комп'ютерного бачення та інших областях.

Архітектура трансформерів базується на механізмах уваги (attention mechanisms) і складається з декількох компонентів:

Енкодер – стек повністю зв'язаних шарів, які приймають на вхід послідовність векторів (наприклад, послідовність слів у реченні) і обробляють її для витягання корисних ознак. Кожен шар енкодера містить два підшари: механізм уваги та нейронну мережу з прямим розповсюдженням.

Декодер – містить стек повністю зв'язаних шарів, але додатково включає механізм уваги, який дозволяє декодеру звертатися до вихідних значень енкодера при генерації послідовностей.

Механізми уваги (Attention Mechanisms) – це ключова складова частина трансформера. Вони дозволяють моделі приділяти різну увагу різним частинам вхідної послідовності під час обробки даних. Зокрема, механізм уваги допомагає моделі зосередитися на важливих елементах вхідних даних, використовуючи ваги, які вивчаються під час навчання. На рисунку 3.2 показано приклад архітектури цієї нейронної мережі.

Механізми уваги трансофрмерів дозволяють моделі приділяти різну увагу різним частинам вхідної послідовності під час обробки даних, а це може бути використано для видалення чутливої інформації.

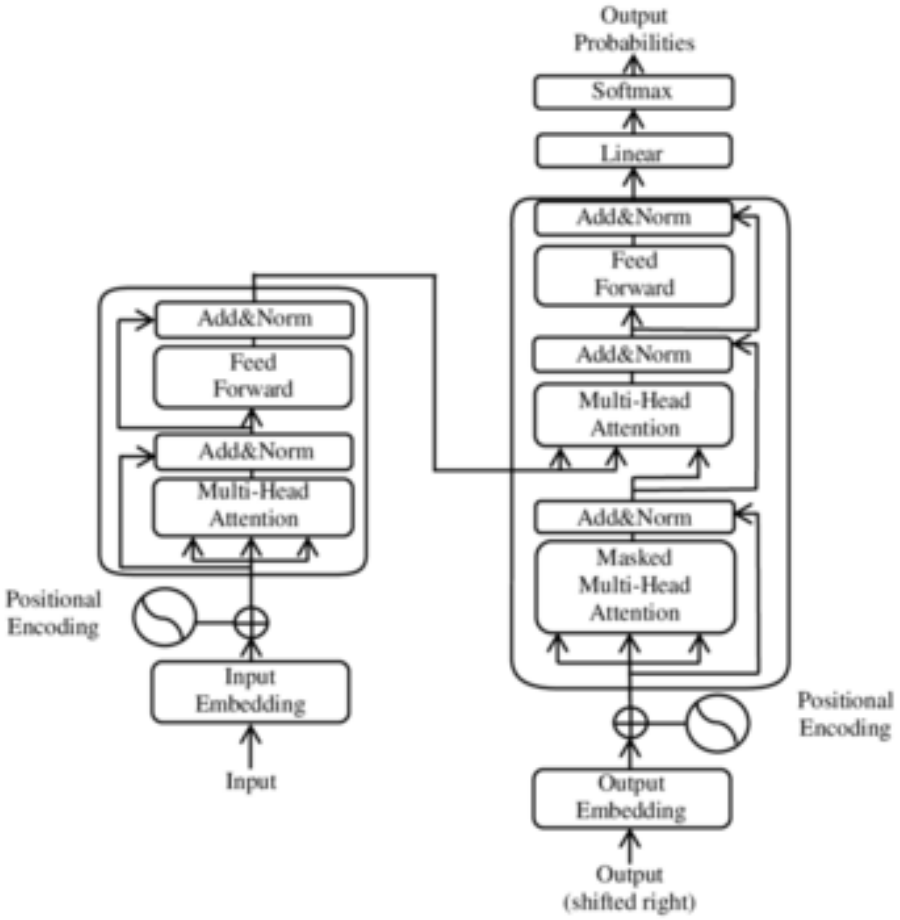

Рисунок 3.2 – Приклад архітектури трансформера

У контексті анонімізації, трансформери можуть бути використані для заміни конкретних імен, адрес або інших особистих даних у тексті або інших послідовностях, замінивши їх загальними токенами або символами. Вони можуть аналізувати контекст і важливість кожного слова чи фрази, що дозволяє ефективно заміняти інформацію, зберігаючи при цьому загальну структуру та семантику тексту.

Крім того, трансформери можуть використовуватися для генерації синтетичних даних, які зберігають статистичні характеристики оригінальних даних, але не містять конкретної особистої інформації. Це може бути корисною стратегією для збереження конфіденційності даних під час аналізу та обробки.

3.6 Порівняння GAN та трансофрмерів

Обидва – GAN і трансформери можуть бути використані для анонімізації даних та створення синтетичних даних, але вони мають свої власні переваги та недоліки.

Переваги та недоліки GAN наведені нижче.

Переваги:

– GAN відомі своєю здатністю генерувати високоякісні та реалістичні дані, що допомагає зберегти структуру та корисну інформацію під час анонімізації даних;

– GAN можуть працювати з великими обсягами даних та генерувати нові дані швидко та ефективно;

– після навчання моделі GAN процес генерації синтетичних даних може бути автоматизованим, що дозволяє легко масштабувати його для великих наборів даних.

Недоліки:

– навчання GAN може вимагати значних обчислювальних ресурсів та часу, особливо для складних завдань та великих обсягів даних.

– деякі варіанти GAN можуть бути схильні до нестабільності під час навчання, що може вимагати додаткової уваги та налаштувань.

Переваги та недоліки трансформерів наведені нижче.

Переваги:

– трансформери є дуже ефективними у роботі з послідовностями даних, такими як текст або звук, що робить їх ідеальними для анонімізації текстових даних;

– трансформери можуть бути легко налаштовані та адаптовані для різних завдань, що дозволяє їх використання в різних областях;

– механізми уваги, що використовуються в трансформерах, дозволяють моделі враховувати контекст при анонімізації даних.

Недоліки:

– трансформери можуть бути вимогливими до обчислювальних ресурсів, особливо при обробці великих обсягів даних або великих моделей.

– трансформери можуть вимагати великої кількості данних для ефективного навчання, що може бути проблематичним для обробки обмежених наборів даних.

Отже, обидва підходи мають свої переваги та недоліки, і вибір між ними залежить від конкретних потреб та обмежень проекту. У випадку, коли важлива реалістичність та швидкість генерації даних, GAN може бути перевагою. Трансформери вимагають використання великих потужностей, та наборів даних, тому їх доцільно використовувати, коли інші легші відомі способи не мають змоги впоратись з поставленою задачею.

3.7 medGAN

GAN проявляє вражаючу продуктивність у створенні дійсних неперервних даних, але вона має обмеження у створенні дискретних даних. Головна причина полягає в тому, що GAN не вдається вивчити розподіл дискретних даних у їхній оригінальній формі під час процесу оновлення градієнта під час навчання. Для подолання цього обмеження Чой та ін. запропонували інноваційний підхід, який називається медичним GAN (medGAN) для синтезу дискретних медичних записів.

Вони включили автокодер [19] у вихідний GAN, щоб вивчити розподіл дискретних даних. Крім того, вони включили метод міні-пакетного усереднення у адверсальну систему, щоб запобігти проблемі «згортання режиму», з якою зіштовхується GAN, коли він має тенденцію генерувати дані з низькою різноманітністю (рисунок 3.3).

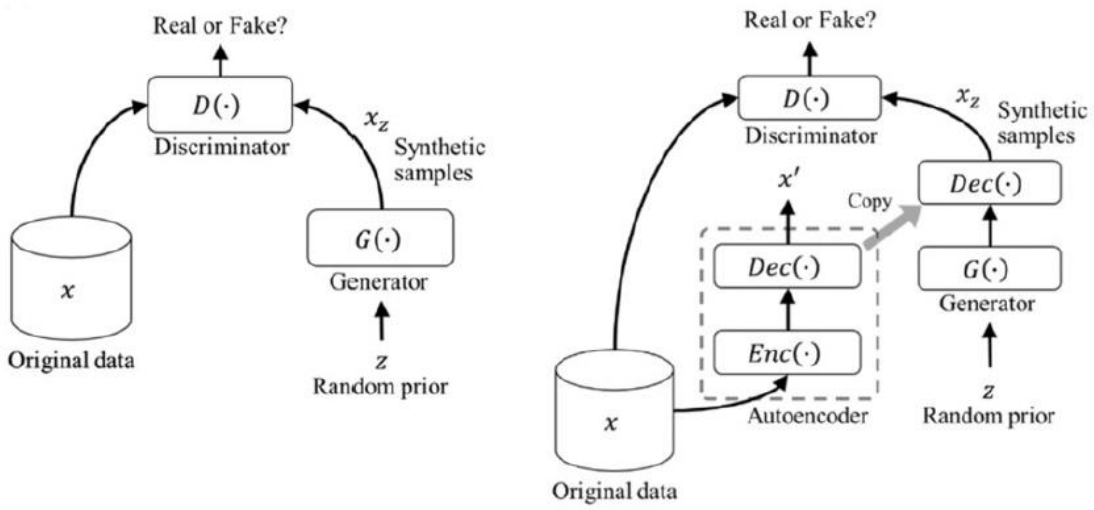

Рисунок 3.3 – Порівняльна характеристика архітектур GAN та medGAN

У межах медичного сектору фреймворк medGAN спрямований на об'єднані дискретні характеристики пацієнтів (наприклад, бінарні та кількісні характеристики), отримані з тривалих медичних записів для проведення експериментів з завданнями машинного навчання. Автори [20] досягли результатів, що не відрізнялися від реальних даних у багатьох експериментах, включаючи статистику розподілу та завдання передбачення.

4 ПРОГРАМНА РЕАЛІЗАЦІЯ ГЕНЕРУВАННЯ СИНТЕТИЧНИХ ДАНИХ

4.1 Огляд засобів розробки

На кожному кроці розробки програмного продукту стоїть велике завдання – вибрати найбільш оптимальні технології, щоб ефективно реалізувати поставлені завдання. Вибір мови програмування стає однією з найважливіших складових, з якими зіштовхуються розробники ще на початковому етапі.

Мова програмування Python виявляється надзвичайно популярною в сферах аналізу даних, машинного навчання та веб-розробки, що робить її відмінним вибором для реалізації різних проектів.

З моменту свого створення Python залишається динамічною та інтерпретованою мовою програмування, яку можна використовувати в різних стилях програмування, включаючи об’єктно-орієнтований та процедурний. Простий синтаксис Python, який нагадує повсякденну англійську мову, робить процес вивчення, розуміння та написання програм більш доступним та приємним.

Python є платформонезалежною мовою програмування, що дозволяє легко переносити код між різними операційними системами, такими як Linux, Windows, MacOS і UNIX, без необхідності внесення будь-яких змін. Крім того, вона підтримує широкий вибір бібліотек, фреймворків та модулів, що спрощують процес розробки, тестування та впровадження програм штучного інтелекту. Серед найпопулярніших бібліотек можна відзначити NumPy, Pandas, SciPy, TensorFlow, PyTorch, scikit-learn і Keras, які значно полегшують написання коду.

Значний обсяг спільноти користувачів сприяє обміну досвідом та знаннями, роблячи Python ще більш привабливим для використання. У 2024

році індекс TIOBE оголосив Python найпопулярнішою мовою програмування у світі [21], що лише підкреслює всі переваги цієї мови.

4.2 Аналіз наборів даних

Датасети, що використовувалися у цьому дослідженні, були отримані з двох джерел. Перше джерело – це база даних Medical Information Mart for Intensive Care (MIMIC-III), яка є вільно доступною загальнодоступною базою даних, що складається з деідентифікованих ЕОМП, пов'язаних з приблизно 60 тис. госпіталізацій пацієнтів у критичних відділеннях Медичного центру Бет Ізраїля Діаконського медичного центру. MIMIC-III містить різні типи медичних даних, серед яких було використано дані діагнозів пацієнтів (DIAGNOSES_ICD) та процедур (PROCEDURES_ICD), закодовані за допомогою Міжнародної статистичної класифікації хвороб та суміжних проблем зі здоров'ям.

У цьому дослідженні було вивчено два різних набори даних MIMIC-III: один набір даних складався з даних про діагнози, а інший (розширений MIMIC-III) складався як з даних про діагнози, так і з даних про процедури. Другим джерелом була Національна база даних з дослідження медичного страхування Тайваню (NHIRD), яка містить дані як про пацієнтів, так і про медичні заклади в рамках програми загального медичного страхування. Доступ до цього набору даних NHIRD обмежений, але дозволяється використовувати його для наукових досліджень. Аналогічно MIMIC-III, ми виділили дані діагнозів пацієнтів, закодовані за допомогою системи МСХ. Важливо зауважити, що хоча наші набори даних стосуються діагнозів і процедур пацієнтів, вони містять багатий набір інформації про різні захворювання, травми, вроджені вади, симптоми, ознаки, аномальні стани, деякі додаткові фактори, що впливають на стан здоров'я, операції та медичні послуги та інше.

Для чесного порівняння з medGAN [22] було зменшено коди МСХ до трьоцифрових кодів для кожного набору даних. Варто зауважити, що у відомостях про продовження ЕОПП для кожного пацієнта (MIMICIII та NHIRD) або про діагнози та процедури (розширений MIMIC-III) кожен рядок відповідає запису про госпіталізацію пацієнта, представлений кодами МСХ. Пацієнт, ймовірно, відвідує лікарню не один раз, тому він може мати кілька записів у відомостях ЕОПП. Було агреговано кожний продовжений запис про пацієнта в один фіксований вектор кодів МСХ. Таким чином, було представлено кожний набір даних у вигляді багатовимірної матриці, де рядок відповідає запису пацієнта, а стовпець – конкретному коду МСХ (наприклад, коду діагнозу або коду процедури). Оскільки коди МСХ агреговані за пацієнтами, вони всі є змінними рахунку. Змінні рахунку вказують кількість разів, коли пацієнт був пов'язаний з конкретним кодом МСХ. У таблиці 4.1 показано частину вибірки рахункового набору даних. Усі значення в таблиці 1 анонімізовані.

Таблиця 4.1 – Частина вибірки даних з лічильного набору

Patient ID	ICD_817	ICD_819	ICD_363
AAAAAA	2	4	5
BBBBBB	0	0	0
CCCCCC	3	2	0
. . . . .	. . .	. . .	. . .
XXXXXX	1	0	4

Зважаючи на те, що всі ознаки в обраних трьох наборах даних, MIMIC-III, розширений MIMIC-III та NHIRD, є лічильними змінними, а на меті маємо проаналізувати як лічильні, так і бінарні дискретні змінні, то було підготовлено бінарну версію кожного лічильного набору даних, перетворивши агреговані лічильні змінні (скажімо, c_i) на бінарні

змінні (скажімо, b_i), використовуючи таке рівняння 4.1. Це дозволяє розширити аналіз до врахування обох типів дискретних змінних та дослідити їх вплив на різноманітні аспекти досліджуваних даних.

$$b_i = \begin{cases} 1, if\ c_i > 0 \\ 0, otherwise \end{cases} \quad (4.1)$$

Таблиця 4.2 показує частину вибірки бінарного набору даних, отриманого з відповідного лічильного набору даних у таблиці 4.1.

Таблиця 4.2 – Частина вибірки даних з бінарного набору

Patient ID	ICD_817	ICD_819	ICD_363
AAAAAA	1	1	1
BBBBBB	0	0	0
CCCCCC	1	1	0
. . . . .	. . .	. . .	. . .
XXXXXX	1	0	1

Бінарні змінні вказують, чи був пацієнт пов'язаний з конкретним кодом МСХ, де кожен код може вказувати на певний медичний стан чи процедуру.

Статистика трьох обраних наборів даних представлена в таблиці 4.3. Набір даних NHIRD є більшим за розміром, ніж набори даних MIMIC-III, враховуючи кількість пацієнтів/записів. У наборі даних MIMIC-III для діагнозів є 942 коди МСХ, 1651 код (діагнозів: 940 та процедурних кодів: 711) у розширеному наборі даних MIMIC-III, і 1015 кодів МСХ в наборі даних NHIRD для діагнозів. Така різноманітність кодів МСХ в різних наборах даних відображає різноманітність медичних умов та процедур, що представлені в цих наборах.

Таблиця 4.3 – Статистика датасетів

MIMIC-III (diagnoses data)	Extended MIMIC-III (diagnoses + procedures data)	NHIRD, Taiwan (diagnoses data)
46517	42214	498909
942	1651	1015
13.99	20.17	8.42
540	640	687
1	2	1

На рисунку 4.1 побудовано ЕФР кількості унікальних кодів МСХ, пов'язаних з усіма пацієнтами у кожному наборі даних.

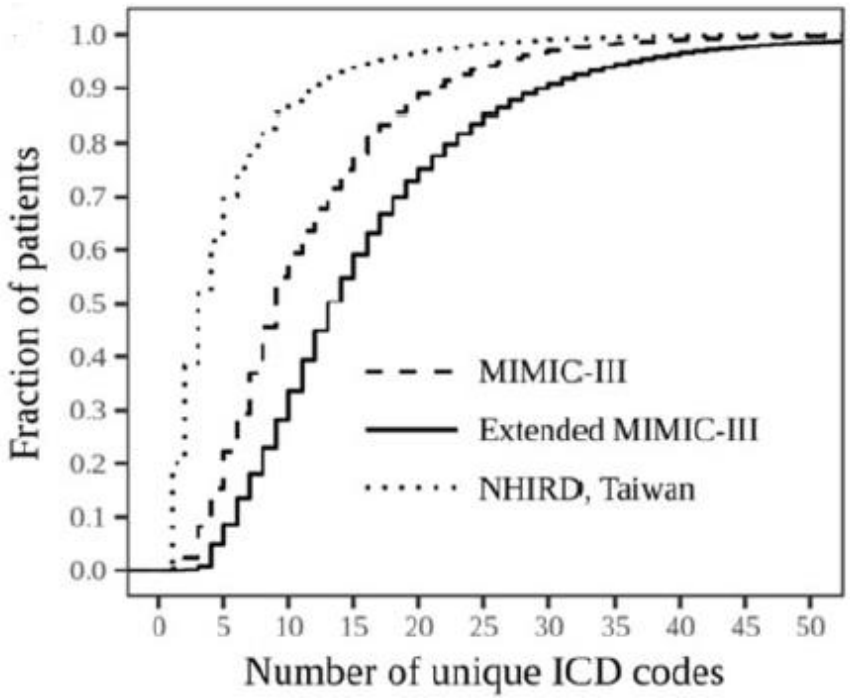

Рисунок 4.1 – Емпірична фунція розподілу кількості унікальних кодів

У NHIRD 70% пацієнтів мають 5 або менше унікальних кодів МСХ, тоді як у MIMIC-III та розширеному MIMIC-III такий же відсоток пацієнтів має до 13 та 18 унікальних кодів МСХ відповідно.

На рисунку 4.2 обчислено відсоток пацієнтів, які пов'язані з кожним кодом МСХ, і потім побудували ЕФР цього відсотка пацієнтів. У NHIRD 90% кодів ICD (913 з 1015) пов'язані лише з 1,31% пацієнтів або менше, тоді як у MIMIC-III 90% кодів ICD (845 з 942) пов'язані з до 2,95% пацієнтів, а у розширеному MIMIC-III 90% кодів ICD (1487 з 1651) пов'язані з до 2,17% пацієнтів.

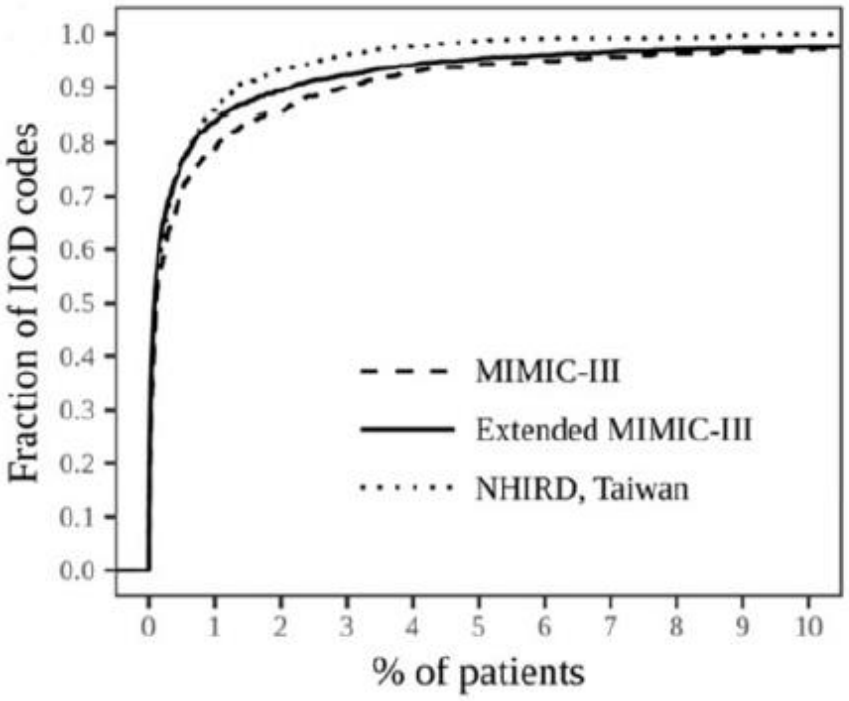

Рисунок 4.2 – Емпірична фунція розподілу відсотка пацієнтів

У таблиці 4.4 наведено 10 найпоширеніших кодів МСХ разом з їх значенням, частотою входжень, кількістю унікальних пацієнтів та відсотком пацієнтів, пов'язаних з кожним кодом у наборі даних діагнозів MIMIC-III. Ці дані допомагають у встановленні важливості кожного коду МСХ та його поширеності серед пацієнтів, що дозволяє здійснювати аналіз та висновки щодо характеристик популяції.

Таблиця 4.4 – Найпоширеніші коди МСХ у датасеті MIMIC-III

Top ICD codes	Meaning	Frequency	No. of patients associated with	Percent of patients associated with
ICD_401	Essential hypertension	21329	17031	38.76%
ICD_427	Cardiac dysrhythmias	20998	14022	30.14%
ICD_428	Heart failure	20676	10154	21.83%
ICD_276	Disorders of fluid, electrolyte, and acid-base balance	20440	12645	27.18%
ICD_250	Diabetes mellitus	16454	10318	22.18%
ICD_414	Other forms of chronic ischemic heart disease	15759	11926	25.64%
ICD_272	Disorders of lipoid metabolism	14768	12268	26.37%
ICD_518	Other diseases of lung	14608	11363	24.43%
ICD_285	Other and unspecified anemias	12910	10631	22.85%
ICD_584	Acute renal failure	11467	9536	20.50%

У таблиці 4.5 наведено 10 найпоширеніших кодів МСХ разом з їх значенням, частотою входжень, кількістю унікальних пацієнтів та відсотком пацієнтів, пов'язаних з кожним кодом у наборі даних діагнозів NHIRD, Taiwan.

Таблиця 4.5 – Найпоширеніші коди МСХ у датасеті NHIRD, Taiwan

Top ICD codes	Meaning	Frequency	No. of patients associated with	Percent of patients associated with
1	2	3	4	5
ICD_250	Diabetes mellitus	170162	44284	8.88%
ICD_401	Essential hypertension	144662	66258	13.28%
ICD_599	Other disorders of urethra and urinary tract	89524	47394	9.50%

Продовження таблиці 4.5

1	2	3	4	5
ICD_295	Schizophrenic disorders	84584	4622	0.93%
ICD_486	Pneumonia, organism unspecified	68484	41982	8.41%
ICD_650	Normal delivery	67437	47154	9.45%
ICD_276	Disorders of fluid, electrolyte, and acid-base balance	66082	42940	8.61%
ICD_414	Other forms of chronic ischemic heart disease	61985	28228	5.66%
ICD_V27	Outcome of delivery	60200	43896	8.80%
ICD_571	Chronic liver disease and cirrhosis	59547	24796	4.97%

У таблиці 4.6 показано десять головних даних пацієнтів для наборів даних MIMIC-III та NHIRD. Основні дані пацієнтів включають частоту (тобто загальну кількість кодів МСХ), загальну кількість унікальних кодів МСХ та відсоток унікальних кодів МСХ для кожного пацієнта.

Таблиця 4.6 – Основні дані пацієнтів

SN. of top patients	MIMIC-III			NHIRD, Taiwan		
	Frequency (No. of total ICD codes)	No. of unique ICD codes	Percent of unique ICD codes	Frequency (No. of total ICD codes)	No. of unique ICD codes	Percent of unique ICD codes
1	2	3	4	5	6	7
1	540	88	9.34%	687	18	1.77%
2	362	85	9.02%	605	5	0.49%
3	361	44	4.67%	527	23	2.27%
4	360	70	7.43%	505	16	1.58%
5	359	61	6.48%	501	5	0.49%
6	332	74	7.86%	490	14	1.38%

Продовження таблиці 4.6

1	2	3	4	5	6	7
7	326	79	8.39%	487	15	1.48%
8	323	42	4.46%	485	20	1.97%
9	316	77	8.17%	469	7	0.63%
10	293	64	6.79%	466	8	0.79%

4.3 Покращені моделі

Модель medGAN показала значні успіхи у генерації синтетичних даних ЕОМП, створюючи реалістичні зразки для лічильних та бінарних змінних. Однак, її продуктивність все ще обмежена в деяких випадках, таких як генерація рідкісних або малопоширених даних, де вона не завжди досягає високої точності. Для подолання цих обмежень були розроблені нові вдосконалені моделі medGAN-W та medGAN-B, які показали кращі результати у тестах, зокрема для лічильних даних та зменшення кількості нульових значень.

Першою запропонованою моделю є medGAN-W, де було використано вдосконалену генеративну мережу під назвою WGAN-GP замість загальної GAN. Решта структури залишається такою ж, як у моделі medGAN. Автори моделі WGAN-GP стверджують, що раніше розроблена модель Вассерштейна (WGAN) допомагає стабілізувати навчання, але генерує низькоякісні зразки або не збігається у деяких налаштуваннях через використання техніки обмеження ваг. Щоб подолати ці проблеми, вони запропонували альтернативний метод обмеження ваг, який називається штрафом за градієнт, що передбачає штрафування норми градієнта дискримінатора (критика) щодо його вхідних даних. Модель WGAN-GP працює краще, ніж багато архітектур GAN, включаючи стандартний WGAN. Таким чином, у цьому дослідженні ми гіпотезували, що застосування моделі medWGAN для генерації синтетичних ЕОМП буде

мати кращу продуктивність, ніж застосування оригінальної моделі medGAN.

Другою запропонованаою моделлю є ще одна альтернатива до medGAN, яка досягається за рахунок заміни традиційної GAN новим алгоритмом, що називається BGAN. У цьому новаторському підході генератор навчається наближати цільовий розподіл, який поступово зближується до справжнього розподілу даних під час оптимізації дискримінатора. Ця мета має на меті тренування генератора для створення зразків, що лежать на границі прийняття рішення поточного дискримінатора під час кожного оновлення. Варто відзначити, що цей алгоритм ефективно працює як з дискретними, так і з неперервними змінними, проявляючи якісно кращі рівні продуктивності порівняно з традиційними GAN. Аналогічно medGAN-W, очікується, що medGAN-W проявить високу продуктивність у генерації синтетичних ЕОМП.

4.4 Опис експерименту

Першим кроком було розділення наборів даних MIMIC-III, розширенний MIMIC-III та NHIRD на дві частини, а саме, навчальний та тестовий набори даних, в співвідношенні 4:1. Було використано навчальний набір для навчання моделей та генерації такої ж кількості синтетичних ЕОМП. Тестовий набір залишився для перевірки прогностичних моделей. Більшість параметрів medGAN були збережені в наших моделях. Деякі загальні параметри перелічені в таблиці 4.7.

Таблиця 4.7 – Загальні параметри експерименту

# of training samples of MIMIC-III	37213
# of training samples of extended MIMIC-III	33771
# of training samples of NHIRD	399127
# of epochs to pre-train the autoencoder	100

Продовження таблиці 4.7

# of epochs to train the model	1000
Batch size	1000
Generator size	(128, 128, 128)
Discriminator size	(256, 128, 1)

Другий кроком було розділення тренувальних даних на навчальні та валідаційні піднабори у співвідношенні 9:1. Попередньо було навчено автокодер протягом 100 епох, використовуючи навчальний піднабір, і для кожної епохи було вказано втрати тренування та валідації, які визначалися як бінарне хрест-ентропійне для бінарних змінних та середнє квадратичне відхилення для змінних рахунку. З кривої тренування помітно, що 100 епох вистачає, і перенавчання відсутнє. Після попереднього навчання автокодера ми копіюємо частину декодера та з'єднуємо її як останній шар генератора G, і навчаємо мережі GAN протягом 1000 епох, використовуючи 90% навчального піднабору. Для кожної епохи було використано решту 10% валідаційного піднабору для перевірки продуктивності (точності та AUC) дискримінатора D як бінарного класифікатора. Ще важливіше використовувати генератор G для випадкової генерації синтетичних даних кожні 10 епох під час тренування та проведення певних перевірок на цих тимчасово згенерованих даних, такі як середні за розмірами і кількість ненульових розмірів. В процесі того як триває процес тренування було помічено, що якість тимчасово згенерованих синтетичних даних стає краще і краще, і всі перевірочні пункти стають стабільними після 700–800 епох у всіх випадках. Було досліджено різні кількості циклів навчання дискримінатора та генератора, які визначено як співвідношення між дискримінатором та генератором, щоб оновлювати їх для кожної епохи тренування. Основуючись на кореляційних коефіцієнтах між середніми за розмірами навчальних даних та кінцевими синтетичними даними, було

встановлено це співвідношення на 2:1 для medGAN та medGAN-W та на 5:1 для medGAN-B.

Навчання моделей та генерування синтетичних даних з розмірами, які були найближчими кратними розміру пакета у навчальних вибірках, як зазначено в таблиці 4.3, тобто 37000, 33000 та 399000 зразків з MIMIC-III, розширений MIMIC-III та NHIRD, відповідно. Спочатку значення сирого згенерованого даних були неперервними у діапазоні від 0 до 1. Їх було конвертовано у бінарний (0 або 1) шляхом округлення.

Подібно до бінарних зразків, для рахункових змінних було використано таку ж кількість навчальних вибірок для навчання моделей та генерації синтетичних даних. Однак сирі значення згенерованих даних були будь-якими неперервними не від'ємними числами. Для цього було округлено неперервні значення синтетичних даних до найближчих цілих значень.

Середній час роботи, потрібний для тренування моделей та генерації синтетичних даних, склав 2 годин для MIMIC-III, 3,5 годин для extended MIMIC-III та 21 година для наборів даних NHIRD.

4.5 Оцінка експерименту

Після генерації синтетичних ЕОМП виникала очевидна задача оцінки цих згенерованих даних і порівняння їх з реальними ЕОМП. Для цих цілей Було використано деякі методи оцінки з двох різних перспектив наступним чином.

Статистичні методи. Як базову перевірку адекватності того, чи навчилися наші моделі розподілу кожного виміру, було розраховано ймовірність для кожного виміру бінарних даних та середнє значення кожного виміру лічильних даних та проведено тест Колмогорова-Смірнова (K–S тест) [23].

Ймовірність за виміром відноситься до ймовірності успіху Бернуллі для кожного виміру (хвороби або процедурного коду) в бінарному наборі даних. Ймовірність за виміром розраховується за допомогою наступної формули:

$$\text{Dimension} - \text{wise probability} = \frac{Number\ of\ patients\ who\ had\ the\ disease\ or\ procedure}{Total\ number\ of\ patients}. \quad (4.2)$$

Середнє значення за виміром відноситься до середнього значення стовпця кожного виміру (хвороби або процедурного коду) у наборі даних лічильників. Середнє значення за виміром обчислюється за допомогою наступної формули:

$$\text{Dimension} - \text{wise average} = \frac{\text{Column sum}}{Total\ number\ of\ records}. \quad (4.3)$$

Було проведено K-S тест для двох вибірок даних (синтетичних даних та реальних даних), щоб перевірити, чи походять обидві вибірки даних з одного розподілу. У K-S тесті статистика обчислюється за допомогою знаходження максимального абсолютного значення відмінностей між кумулятивними функціями розподілу двох вибірок. Нульова гіпотеза полягає в тому, що обидві вибірки походять з однакового розподілу. У цьому експерименті було відкинуто нульову гіпотезу з низьким значенням (зазвичай < 0,05).

Методи машинного навчання. Було використано видобуток асоціативних правил та прогнозування за вимірами для тестування того, як зберігаються міжвимірні відносини в синтетичних даних.

Видобуток асоціативних правил, таких як Apriori, широко використовуються на даних ЕОМП для виявлення асоціацій та зрозумілих зв'язків між клінічними концепціями (лікарські засоби, результати лабораторних досліджень та діагнози проблем) [24], [25]. Цей метод

машинного навчання використовувався на основі правил для виявлення деяких сильних асоціацій або зв'язків між змінними як у реальних, так і в синтетичних наборах даних. Перевірялось чи знайдені взаємозв'язки у реальному наборі даних присутні в відповідному синтетичному наборі даних.

Для простоти було розглянуто лише однозначні відносини, при цьому всі правила мали довжину 2. Для MIMIC-III та NHIRD було встановлено параметри алгоритму Apriori (пороги підтримки та достовірності) так, щоб значення наближалося до 50–200 правил у реальному наборі даних, і використано ті ж параметри для синтетичних наборів даних, згенерованих за допомогою трьох моделей GAN. Для порівняння правил, знайдених у кожному з реальних наборів даних, з правилами, знайденими у відповідному синтетичному наборі даних, використовувалось декілька метрик, таких як точність і повнота.

Прогнозування за вимірами як опосередкований засіб для тестування міжвимірних відносин у синтетичних даних, було використано завдання прогнозування для кожного коду МСХ. Було застосовано три популярні методи машинного навчання – логістичну регресію, випадковий ліс та метод опорних векторів (SVM), які зазвичай використовуються для прогнозування на даних ЕОМП [26].

Результати прогнозування порівнювались за вимірами прогностичних моделей, навчених на синтетичних даних, з відповідними реальними даними.

Для опису більш конкретно, припустимо, що у нас є загалом n вимірів (хвороби або процедурні коди), де n = 942 для MIMIC-III, n = 1651 для розширеного MIMIC-III та n = 1015 для NHIRD. Прогностичний алгоритм розглядає вимір $y \in n$ як міру цільової або залежної змінної для прогнозування (тобто, чи може статися ця хвороба або процедура), а решта (n – 1) вимірів як ознаки або незалежні змінні x.

Варто зазначити, що для набору даних лічильника цільова змінна у перетворюється в бінарну за допомогою тієї ж техніки, що й у формулі 4.1. Таким чином, було побудовано прогностичні моделі для кожної хвороби або процедури, використовуючи як реальні, так і синтетичні набори даних, і ці моделі подальше застосовувались до тестових даних для отримання показників ефективності (F1-показники). Для порівняння результатів прогнозування реальних та синтетичних наборів даних було обчислено коефіцієнти кореляції (CC) та середньоквадратичні помилки (RMSE) з використанням всіх F1-показників по всіх вимірах.

4.6 Результати експерименту

Було застосувано 3 різні генеративні моделі, а саме, існуюча medGAN та дві нові запрограмовані можелі на її основі, а саме – medGAN-W та medGAN-B, окремо до наборів даних для генерації синтетичних даних.

На рисунках 4.2, 4.3, 4.4 показано продуктивність ймовірності по вимірам для трьох різних генеративних моделей для синтетичних бінарних даних MIMIC-III, розширеного MIMIC-III та NHIRD.

Кожен розсіюваний графік відображає продуктивність однієї генеративної моделі. На розсіюваних графіках кожна точка представляє один код MCX. Вісь x відображає ймовірність успіху Бернуллі кожної хвороби або процедури (код MCX) у реальних даних, а вісь y відображає ймовірність успіху кожної хвороби або процедури (код MCX) у синтетичних даних. Діагональна лінія показує ідеальний випадок, коли продуктивність синтетичних даних ідентична реальним даним.

Щоб чисельно виміряти продуктивність кожної генеративної моделі, було використано коефіцієнти кореляції та середньоквадратичні помилки між реальними та синтетичними даними.

Графіки на рисунках 4.2, 4.3, 4.4 показують схожі тенденції ймовірності по вимірам для MIMIC-III, розширеного MIMIC-III та NHIRD

бінарних даних. Запропоновані моделі medGAN-W та medGAN-B демонструють трохи кращу продуктивність порівняно з базовою моделлю medGAN, але рівні продуктивності дуже близькі до найвищих показників (100%). Серед трьох моделей medGAN-B має найкращу продуктивність.

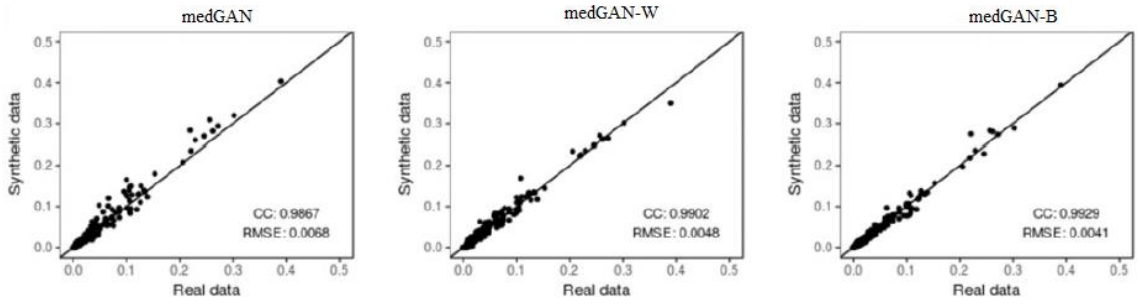

Рисунок 4.2 – Продуктивність ймовірності по вимірам датасету MIMIC-III

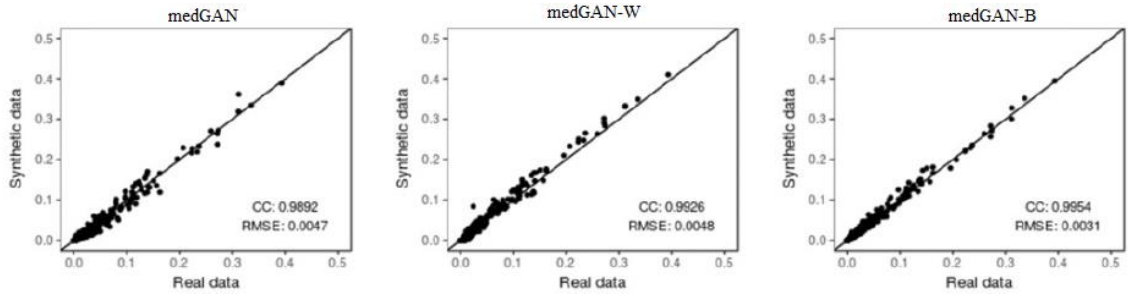

Рисунок 4.3 – Продуктивність ймовірності по вимірам датасету MIMIC-III (розширений)

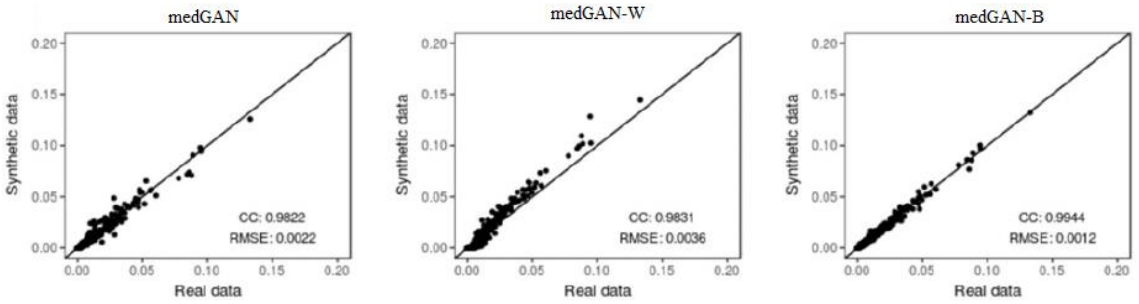

Рисунок 4.4 – Продуктивність ймовірності по вимірам датасету NHIRD

На рисунках 4.5, 4.6, 4.7 показано середню кількість по вимірам для трьох різних генеративних моделей для синтетичних лічильничих даних MIMIC-III, розширеного MIMIC-III та NHIRD.

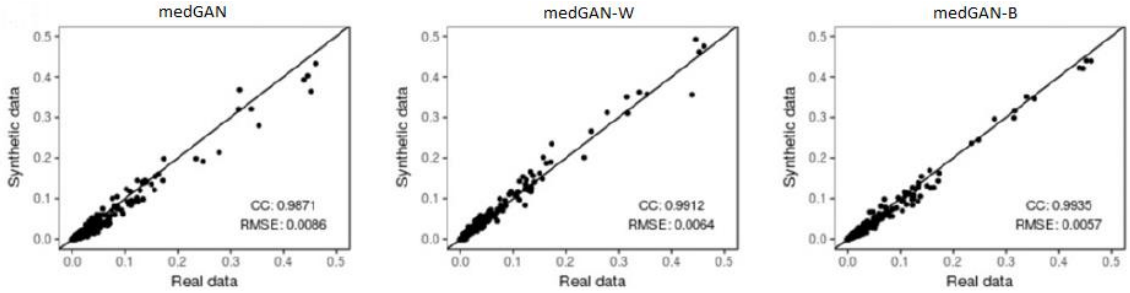

Рисунок 4.5 – Середня кількість по вимірам датасету MIMIC-III

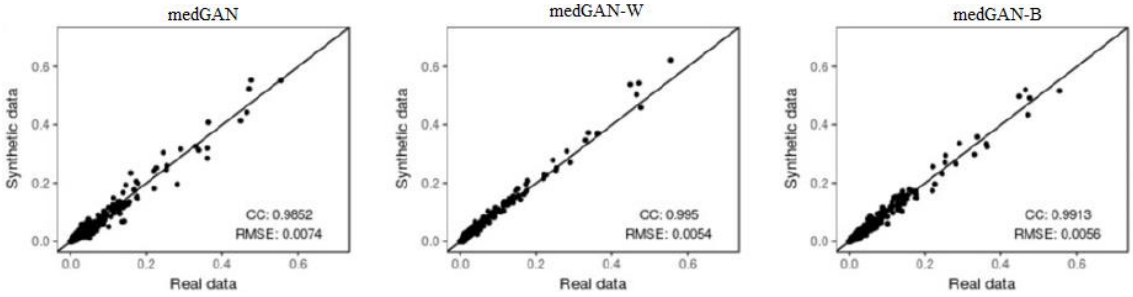

Рисунок 4.6 – Середня кількість по вимірам датасету MIMIC-III (розширений)

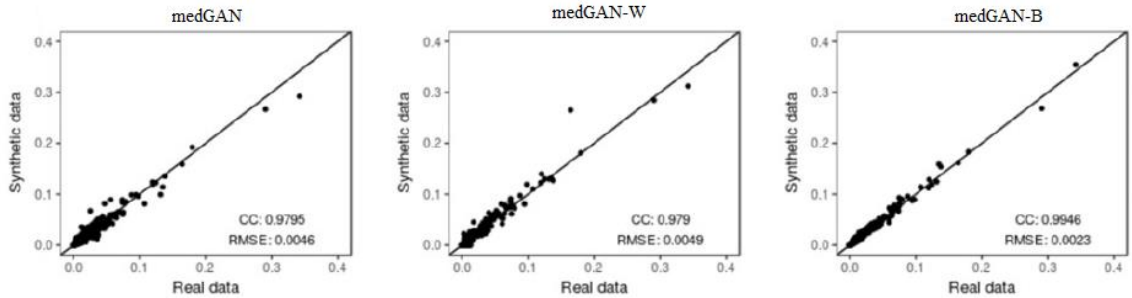

Рисунок 4.7 – Середня кількість по вимірам датасету NHIRD

Кожний розсіюваний графік відображає продуктивність однієї генеративної моделі. На розсіюваних графіках кожна точка представляє 1 код МСХ. Вісь x відображає середню кількість кожної хвороби або процедури (код МСХ) у реальних даних, а вісь y відображає середню кількість кожної хвороби або процедури (код МСХ) у синтетичних даних. Згідно з рисунками 4.5 та 4.6 для лічильничих даних MIMIC-III та розширеного MIMIC-III обидві моделі medGAN-W та medGAN-B показують невелике покращення порівняно з medGAN. Однак для лічильничих даних NHIRD на рисунку 4.7 лише medGAN-B перевершує medGAN, а результати medGAN та medGAN-W при цьому майже ідентичні.

Було застосовано тест Колмогорова-Смірнова (K-S) по вимірам, щоб перевірити, чи певний зразок (наприклад, xi) синтетичних даних та відповідний зразок (наприклад, yi) реальних даних з тією самою назвою вимірюється з однаковим розподілом (1 або 0). Потім обчислювався загальний відсоток подібності між кожним синтетичним набором даних та відповідним реальним набором даних. Отримані результати узагальнені у таблиці 4.8. На основі результатів тесту Колмогорова–Смірнова можна зробити висновок про подібність розподілів синтетичних та реальних даних для кожної моделі. Це дозволяє оцінити ефективність генеративних моделей у відтворенні реальних розподілів даних.

Таблиця 4.8 – Відсоток подібності для всіх методів

Dataset	Data type	Model	K-S test similarity
1	2	3	4
MIMIC-III	Binary	medGAN	94,48%
		medGAN-W	95,97%
		medGAN-B	**97,45%**
	Count	medGAN	88,64%
		medGAN-W	**95,12%**
		medGAN-B	89,70%

Продовження таблиці 4.8

1	2	3	4
Extended MIMIC-III	Binary	medGAN	95,34%
		medGAN-W	96,49%
		medGAN-B	**97,64%**
	Count	medGAN	93,46%
		medGAN-W	**96,24%**
		medGAN-B	94,12%
NHIRD, Taiwan	Binary	medGAN	92,12%
		medGAN-W	76,35%
		medGAN-B	**95,86%**
	Count	medGAN	83,35%
		medGAN-W	80,59%
		medGAN-B	**86,31%**

У таблиці 4.8 показано відсоток подібності між синтетичними даними, згенерованими трьома генеративними моделями, та їх реальними аналогами даних. Зауважимо, що подібно до попередніх статистичних результатів, запропоновані medGAN-W та medGAN-B перевершують medGAN. У більшості випадків medGAN-B має найкращу продуктивність. medGAN-W демонструє кращий результат лише для лічильничих даних MIMIC-III та розширеного MIMIC-III.

Основною метою для асоціаційного аналізу правил у цьому дослідженні є перевірка того, як зберігаються міжвимірні зв'язки у синтетичних даних, а не пошук найкращої продуктивності з точки зору знаходження найбільшої кількості правил співвипадковостей у реальних даних. Спробувавши кілька наборів параметрів з мінімальною підтримкою від 5% до 10% та достовірністю від 40% до 50% було отримано майже однакові результати при оцінці міжвимірних зв'язків між реальними даними та синтетичними даними (визначеною за точністю/повторюваністю) у всіх

випадках. Через схожість результатів було вирішено показувати лише один результат оцінки, який виробляв приблизно 50~200 правил з реальних наборів даних.

Було знайдено 72 правила з реального набору даних MIMIC-III та 154 правила з реального набору даних MIMIC-III(розширений), використовуючи алгоритм Apriori, встановивши мінімальну підтримку = 0.05, мінімальну достовірність = 0.5, мінімальну довжину = 2 та максимальну довжину = 2. Щодо NHIRD, який є розрідженим та більшим за MIMIC-III, встановлено мінімальну підтримку = 0.01, мінімальну достовірність = 0.40, мінімальну довжину = 2 та максимальну довжину = 2, і знайдено 63 правила з реального набору даних.

Це налаштування параметрів було збережно для відповідних синтетичних наборів даних. Кількість правил, знайдених у всіх синтетичних наборах даних, а також точності та повторюваності, узагальнені у таблиці 4.9.

Таблиця 4.9 – Результати асоціативного аналізу правил

Dataset	No. of extracted rules in real data	Generative model	No. of extracted rules in synthetic data	No. of matched rules in synthetic data	Precision	Recall
MIMIC-III	72	medGAN	180	61	0.3388	0.8472
		medGAN-W	64	52	0.8125	0.7222
		medGAN-B	153	67	0.4379	0.9305
Extended MIMIC-III	154	medGAN	274	134	0.489	0.8701
		medGAN-W	201	142	0.7064	0.922
		medGAN-B	229	150	0.655	0.974
NHIRD, Taiwan	63	medGAN	1350	56	0.0414	0.8888
		medGAN-W	62	50	0.8064	0.7936
		medGAN-B	520	60	0.1153	0.9523

Як показано у таблиці 4.9, за однакових налаштувань параметрів, medGAN-W дає найвищу точність, а medGAN-B – найвищу повторюваність для набору даних MIMIC-III. У цьому випадку спостерігається, що як запропоновані medGAN-W, так і medGAN-B перевершують оригінальний medGAN.

Аналіз асоціаційних правил на розширеному наборі даних MIMIC-III виводить результати, схожі на MIMIC-III. Датасет NHIRD подібно до MIMIC-III, спостерігається, що medGAN-W дає найвищу точність, а medGAN-B – найвищу повторюваність. Таким чином, можна зробити висновок, що створені моделі перевершують medGAN.

З асоціаційного аналізу правил видно, що medGAN-B може відтворювати більшість правил, які були помічені в реальних даних, і, отже, він виводить найкращу повторюваність (93.05% для MIMIC-III, 97.40% для розширеного MIMIC-III та 95.23% для NHIRD). Натомість medGAN-W генерує найменшу кількість невірних правил у синтетичних даних, і, отже, він виводить найкращу точність (81.25% для MIMIC-III, 70.64% для розширеного MIMIC-III та 80.64% для NHIRD). Хоча medGAN-B показує низьку точність для даних NHIRD, він працює краще, ніж medGAN.

Важливою частиною визначення того, наскільки добре синтетичні дані, створені генеративними моделями, виконуються порівняно з реальними даними є виконання завдань машинного навчання. Тут показується виступ прогнозування по вимірах як для бінарних, так і для лічильних змінних для синтетичних даних MIMIC-III, MIMIC-III (розширений) та NHIRD з використанням кожної з трьох вищезазначених прогностичних моделей.

На рисунках 4.8, 4.9, 4.10 показано ефективність прогнозування по вимірах трьох генеративних моделей, отриманий з результатів моделі логістичної регресії, навчаної на синтетичних бінарних даних MIMIC-III, MIMIC-III(розширений) та NHIRD, та відповідних реальних даних.

На діаграмах розсіювання кожна точка представляє один код МСХ. На вісі x відображаються оцінки F1 моделі логістичної регресії, навчаної на реальних бінарних даних, а на вісі y – оцінки F1 моделі логістичної регресії, навчаної на синтетичних бінарних даних.

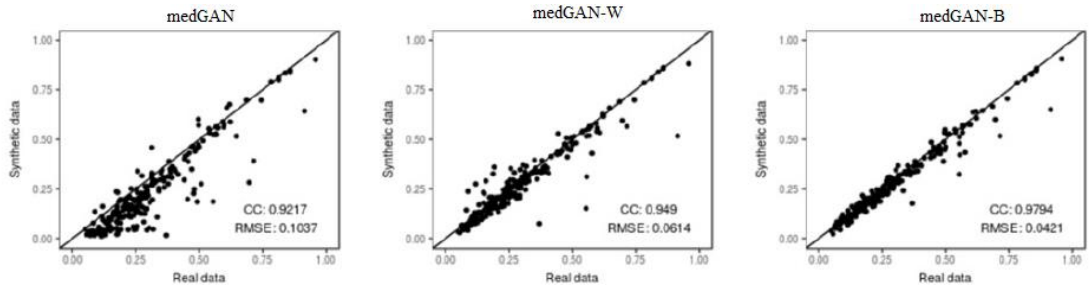

Рисунок 4.8 – Результати прогнозування для бінарних даних датасету MIMIC-III

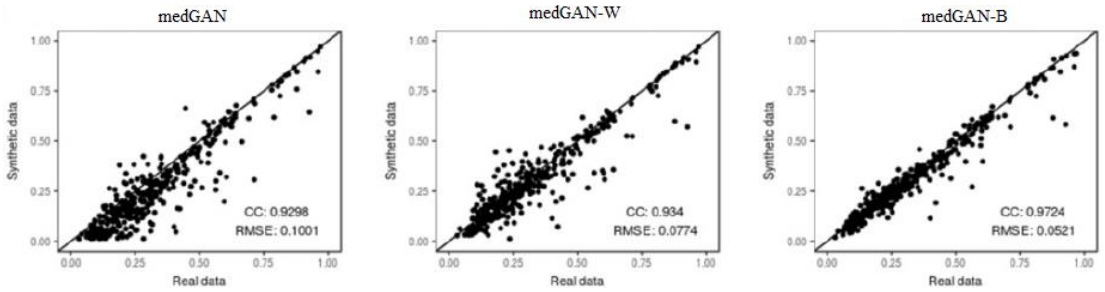

Рисунок 4.9 – Результати прогнозування для бінарних даних датасету MIMIC-III (розширений)

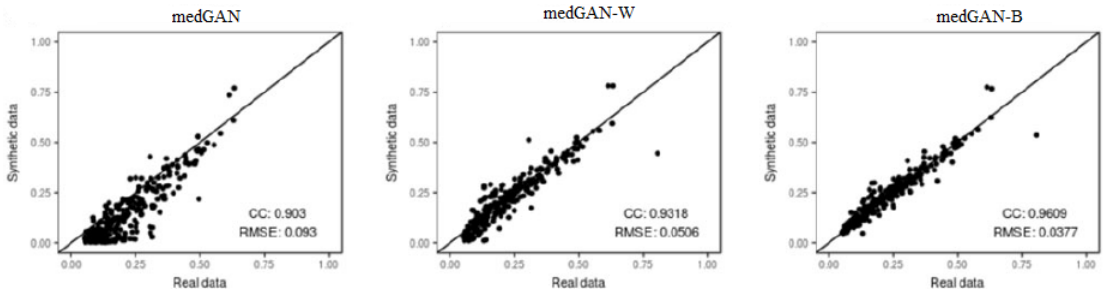

Рисунок 4.10 – Результати прогнозування для бінарних даних датасету NHIRD

Щодо результатів прогнозування для бінарних даних MIMIC-III, MIMIC-III(розширений) та NHIRD, як показано на рисунках 4.8, 4.9, 4.10, як medGAN-W, так і medGAN-B видають кращі результати, ніж medGAN. Особливо варто відзначити, що medGAN-B має найкращу продуктивність серед всіх генеративних моделей.

На рисунках 4.11, 4.12, 4.13 показано ефективність прогнозування по вимірах трьох генеративних моделей, отриманий з результатів моделі логістиної регресії, навчаної на синтетичних лічильних даних MIMIC-III, MIMIC-III (розширений) та NHIRD, та відповідних реальних даних. На діаграмах розсіювання кожна точка представляє один код МСХ. На вісі x відображаються оцінки F1 моделі логістичної регресії, навчаної на реальних лічильних даних, а на вісі y – оцінки F1 моделі логістичної регресії, навчаної на синтетичних лічильних даних. Щодо виступу прогнозування по вимірах для лічильних даних MIMIC-III, розширених MIMIC-III, як показано на рисунках 4.11, 4.12, як medGAN-W, так і medGAN-B видають кращі результати, ніж medGAN, але medGAN-W має найкращий виступ для MIMIC-III, а medGAN-B – для MIMIC-III (розширений), хоча вони дуже близькі у обох випадках. Натомість для лічильних даних NHIRD на рисунку 4.13 medGAN-W має трохи кращий рівень продуктивності, ніж інші моделі, але не спостерігається значних відмінностей серед цих трьох генеративних моделей.

Було оцінено результати прогнозування інших двох класифікаторів машинного навчання, випадкового лісу та методу опорних векторів (SVM), аналогічно тому, як це було зроблено для методу логістичної регресії, який розглядався вище. Результати прогнозування трьох генеративних моделей, отримані з результатів (оцінок F1) трьох прогностичних класифікаторів, показані в таблиці 4.10. У результатах прогнозування випадкового лісу видно, що medGAN-B показує кращі результати, ніж інші дві генеративні моделі, за винятком MIMIC-III (розширений) та лічильних наборів даних NHIRD.

Таблиця 4.10 – Прогнози трьох генеративних методів

Dataset	Data type	Generative model	Logistic regression	Random forest	SVM
MIMIC-III	Binary	medGAN	0.9217	0.8907	0.9406
		medGAN-W	0.949	0.9564	0.9505
		medGAN-B	0.9794	0.9733	0.954
	Count	medGAN	0.905	0.919	0.9469
		medGAN-W	0.9582	0.947	0.9507
		medGAN-B	0.9441	0.9593	0.9589
Extended MIMIC-III	Binary	medGAN	0.9298	0.9248	0.9445
		medGAN-W	0.934	0.945	0.9389
		medGAN-B	0.9724	0.97	0.9655
	Count	medGAN	0.9255	0.8985	0.9278
		medGAN-W	0.9503	0.9371	0.9474
		medGAN-B	0.9615	0.9282	0.9553
NHIRD, Taiwan	Binary	medGAN	0.903	0.8339	0.897
		medGAN-W	0.9318	0.8471	0.9132
		medGAN-B	0.9609	0.9232	0.9705
	Count	medGAN	0.9639	0.9325	0.9705
		medGAN-W	0.9702	0.9325	0.9723
		medGAN-B	0.9656	0.9282	0.9756

У прогнозах методу SVM medGAN-B завжди перевершує інші генеративні моделі, хоча у деяких випадках результати дуже близькі. Таблиця 4.11 узагальнює результати прогнозування, що показує найкращі генеративні моделі завдань прогнозування на різних синтетичних даних.

Таблиця 4.11 – Узагальнення результатів прогнозування

Dataset	Data type	Logistic regression	Random forest	SVM
1	2	3	4	5
MIMIC-III	Binary	medGAN-B	medGAN-B	medGAN-B
	Count	medGAN-W	medGAN-B	medGAN-B

Продовження таблиці 4.11

1	2	3	4	5
Extended MIMIC-III	Binary	medGAN-B	medGAN-B	medGAN-B
	Count	medGAN-B	medGAN-W	medGAN-B
NHIRD, Taiwan	Binary	medGAN-B	medGAN-B	medGAN-B
	Count	medGAN-W	medGAN-W	medGAN-B

З таблиць 4.10 та 4.11 можна сказати, що створені моделі (medGAN-B і medGAN-W) перевершують базову модель medGAN для кожного з трьох завдань прогностичного моделювання.

Підсумки оцінки результатів представлено в таблиці 4.12. Таблиця показує найкращу модель для кожного критерію оцінки синтетичних наборів даних. Очевидно, що в кожному випадку оцінок створені моделі, будь то medGAN-B або medGAN-W, перевершують базову модель medGAN.

Таблиця 4.12 – Підсумкові оцінки результатів

Dataset	Data type	Evaluation criteria	Best generative model
1	2	3	4
MIMIC-III	Binary	Dimension-wise probability performance	medGAN-B /medGAN-W
		K–S test	medGAN-B
		Association rule mining	medGAN-B /medGAN-W
		Dimension-wise prediction performance	medGAN-B
	Count	Dimension-wise average count	medGAN-B /medGAN-W
		K–S test	medGAN-W
		Dimension-wise prediction performance	medGAN-B

Продовження таблиці 4.12

1	2	3	4
Extended MIMIC-III	Binary	Dimension-wise probability performance	medGAN-B /medGAN-W
		K–S test	medGAN-B
		Association rule mining	medGAN-B /medGAN-W
		Dimension-wise prediction performance	medGAN-B
	Count	Dimension-wise average count	medGAN-B /medGAN-W
		K–S test	medGAN-W
		Dimension-wise prediction performance	medGAN-B
NHIRD, Taiwan	Binary	Dimension-wise probability performance	medGAN-B
		K–S test	medGAN-B
		Association rule mining	medGAN-B /medGAN-W
		Dimension-wise prediction performance	medGAN-B
	Count	Dimension-wise average count	medGAN-B
		K–S test	medGAN-B
		Dimension-wise prediction performance	medGAN-W

У дуже небагатьох випадках покращення, запропоновані новими моделями, не були значними. Проте в більшості випадків було отримано вражаючі результати як для бінарних, так і для лічильних даних.

Використання двох різних наборів даних MIMIC-III у цьому дослідженні мало важливу мету – перевірити, чи можуть запропоновані моделі бути застосовані до набору даних, що включає кілька типів МСХ даних одночасно. З цієї причини, на додаток до набору даних діагнозів MIMIC-III, було використано розширений набір даних MIMIC-III, який включав як дані діагнозів, так і процедури МСХ. Таблиця 4.12 показує, що

результати оцінки розширеного набору даних MIMIC-III такі ж, як і для набору даних MIMIC-III, що підтверджує ефективність наших моделей.

Порівняння запропонованих моделей medGAN-W та medGAN-B є виправданим. Для даних MIMIC-III medGAN-W перевершує medGAN-B тільки в тесті K–S для лічильних даних, і medGAN-B показує найкращу продуктивність для всіх інших оцінок. На противагу цьому, в даних NHIRD medGAN-B показує найкращу продуктивність у всіх випадках, крім прогнозування лічильних даних. Проте покращення medGAN-W було незначним. У добуванні асоціативних правил кожна модель показує кращу продуктивність з різних перспектив. Оскільки в кількох випадках medGAN-W показує незначне покращення або співставну продуктивність з medGAN-B, його продуктивність було проаналізувано з іншої точки зору, а саме загальної кількості всіх нульових вимірів у синтетичних даних.

Під час створення синтетичних даних було помічено, що в нашому реальному наборі даних деякі хвороби рідко зустрічалися серед пацієнтів, тобто деякі виміри (стовпці) складалися лише з нулів або дуже малої кількості ненульових значень. Для цих вимірів моделі могли генерувати синтетичні дані з деякими повністю нульовими вимірами. Таблиця 4.13 наводить ці статистичні дані для лічильних наборів даних, показуючи, що для синтетичних наборів даних з лічильними змінними medGAN-W генерує більше вимірів з нулями, ніж medGAN-B і medGAN.

Таблиця 4.13 – Загальна кількість всіх нульових вимірів

Dataset (count variables)	MIMIC-III (total dimensions: 942)	Extended MIMIC-III (total dimensions: 1651)	NHIRD, Taiwan (Total dimensions: 1015)
1	2	3	4
Original (real) data	6 (0.64 %)	26 (1.57 %)	5 (0.49 %)
medGAN synthetic data	324 (34.39 %)	620 (37.55 %)	172 (16.94 %)

Продовження таблиці 4.13

1	2	3	4
medGAN-W synthetic data	607 (64.43 %)	1135 (68.75 %)	694 (68.37 %)
medGAN-B synthetic data	163 (17.30 %)	453 (27.44 %)	128 (12.61 %)

Повністю нульові виміри, створені medGAN-W, становлять 607 (64,43%) для даних MIMIC-III, 1135 (68,75%) для MIMIC-III (розширений) та 694 (68,37%) для NHIRD. Хоча medGAN тут показує гарні результати, як показано в таблиці 4.13, він не продемонстрував кращу продуктивність порівняно з medGAN-W в інших попередніх оцінках. Навпаки, medGAN-B показує найкращі результати та створює меншу кількість повністю нульових вимірів (17,30% для даних MIMIC-III, 27,44% для розширених даних MIMIC-III та 12,61% для наборів даних NHIRD). Тому загалом можна зробити висновок, що запропонований medGAN-B перевершує як medGAN-W, так і medGAN.

Дослідження проводилося для створення реалістичних і корисних дискретних синтетичних даних ЕОМП, використовуючи ідею покращених GAN. У цій роботі, окрім базового статистичного аналізу, було застосовано три популярні методи машинного навчання для прогнозного моделювання та один для добування асоціативних правил. Все дослідження проводилося на трьох різноманітних наборах даних ЕОМП – MIMIC-III (дані діагнозів), розширений MIMIC-III (дані діагнозів та процедур) та NHIRD (дані діагнозів) – з точки зору їхнього джерела, розміру та розрідженості. Результати оцінки всіх проведених експериментів доводять перевагу запропонованих моделей над існуючою моделлю medGAN у створенні реалістичних синтетичних даних ЕОМП. Це також гарний результат, бо є розуміння, що згенеровані синтетичні дані є достатньо якісними для завдань машинного навчання. Важливо зазначити, що в цій роботі досліджувались дані про діагнози та процедури пацієнтів як кейс-стаді.

Проте, запропоновані методи не обмежуються лише цими даними, оскільки не використовувалось жодних знань, специфічних для діагнозів чи процедур, під час навчання GAN. Додатково, оригінальні методи, засновані на GAN, добре працюють для генерації безперервних даних. Тому, як загальний метод, створена модель може використовуватися для створення будь-яких реалістичних даних ЕОМП, навіть поза межами медичної сфери.

Оскільки синтетичні дані створюються штучно, тому немає явного відображення між реальними та синтетичними даними. З цієї причини, можна сказати, що згенеровані синтетичні дані також залишаються стійкими до повторної ідентифікації. Більш важливо, Чой та ін. провели формальну оцінку ризиків конфіденційності medGAN на основі як атрибутивного розкриття, так і розкриття присутності в синтетичному наборі даних. Експерименти з конфіденційності показали, що medGAN генерує різноманітні синтетичні зразки, які розкривають мало інформації потенційним зловмисникам. Оскільки ми використовували архітектуру, подібну до medGAN, вона успадковує збереження конфіденційності, а також перевершує її, оскільки якість даних була покращена.

Було запропоновано дві нові створені моделі на основі medGAN, а саме medGAN-W та medGAN-B, які можуть адекватно навчатися на розподілі реальних ЕОМП і демонструють видатну продуктивність у створенні реалістичних синтетичних ЕОМП для як бінарних, так і лічильних змінних. Всебічно було проаналізовано синтетичні дані ЕОМП, згенеровані трьома генеративними моделями, та порівняно їхні результати оцінки з реальними даними ЕОМП. На основі цього дослідження зроблено висновок, що запропоновані моделі перевершують існуючу модель medGAN, і що серед цих трьох моделей medGAN-B показує найкращі результати.

ВИСНОВКИ

У рамках кваліфікаційної роботи був проведений збір матеріалів, які стосуються теми – «Анонімізація великих даних в інтелектуальних медичних системах».

В ході дослідження було опрацьовано існуючі методи, а також створено покращені моделі, які забезпечать анонімізацію великих медичних даних. Ці методи повинні ефективно замінювати ідентифікатори на неідентифікуючі дані, забезпечуючи при цьому збереження корисної інформації для подальшого використання.

Було розглянуто створення реалістичних і корисних дискретних синтетичних даних електронних медичних записів з використанням покращених генеративних змагальних мереж (GAN). У цій роботі, окрім базового статистичного аналізу, було застосовано три популярні методи машинного навчання для прогнозного моделювання та один метод для добування асоціативних правил.

Було запропоновано дві нові моделі на основі medGAN, які адекватно навчаються на розподілі реальних ЕОМП і демонструють високу продуктивність у створенні реалістичних синтетичних ЕОМП для бінарних та лічильних змінних. Порівняльний аналіз показав, що запропоновані моделі перевершують існуючу модель medGAN у створенні реалістичних синтетичних даних ЕОМП, підтверджуючи, що згенеровані дані є достатньо якісними для завдань машинного навчання.

Було розглянуто важливий аспект – анонімізацію даних. Синтетичні дані створюються штучно, що забезпечує відсутність явного відображення між реальними та синтетичними даними, зменшуючи ризики повторної ідентифікації. Формальна оцінка ризиків конфіденційності показала, що medGAN генерує різноманітні синтетичні зразки, які розкривають мінімум інформації потенційним зловмисникам. Оскільки запропоновані моделі

мають подібну архітектуру, вони успадковують ці властивості конфіденційності та покращують якість даних.

Зроблено висновок, що серед трьох розглянутих моделей створені моделі перевершують існуючу модель medGAN. Це підтверджує ефективність запропонованих підходів у створенні реалістичних синтетичних даних ЕОМП, які можуть бути використані для широкого спектра завдань, включаючи анонімізацію великих даних.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1) CareZone | Easily manage multiple medications and health info. *CareZone | Easily manage multiple medications and health info*. URL: https://carezone.com/ (date of access: 19.03.2024).

2) Limited P. D. U. Pregnancy + | tracker app – додатки в google play. *Android Apps on Google Play*. URL: https://play.google.com/store/apps/details?id=com.hp.pregnancy.lite&hl=uk&gl=US (дата звернення: 19.03.2024).

3) Воронова Д. С. Застосунок для збереження та порівняння результатів медичних аналізів: кваліфікаційна робота бакалавра. Харків: ХНУРЕ, 2022. 61 с.

4) Види медичних аналізів. URL: https://finemed.com.ua/ua/10-articles/laboratoriya/96-analizi-kharkiv-5 (дата звернення 19.03.2024).

5) Про HELSI. helsi.me. URL: https://helsi.me/about (дата звернення: 21.03.2024).

6) Загальний регламент про захист даних (GDPR) – GDPR-Text.com. GDPR-Text.com - GDPR Text, Translation and Commentary. URL: https://gdpr-text.com/uk/ (дата звернення: 01.03.2024).

7) Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*. 2002. №10 (5). P. 571–588.

8) Golle P. Revisiting the uniqueness of simple demographics in the us population. *5th ACM Workshop on Privacy in Electronic Society (WPES)*. 2006. P. 77–80.

9) Hafner K. And if you liked the movie, a Netflix contest may reward you handsomely. New York Times, October 6. 2006.

10) Narayanan A., Shmatikov V. Robust de-anonymization of large sparse datasets. *IEEE Symposium on Security and Privacy (SP)*. 2008. P. 111–125.

11) Barbaro M., Zeller T. A face is exposed for AOL searcher no. 4417749. New York Times. 2006.

12) Data anonymization for pervasive healthcare: a systematic mapping study (preprint) / N. Al Moubayed et al. *JMIR Medical Informatics*. 2021. URL: https://doi.org/10.2196/29871 (date of access: 21.03.2024).

13) Воронова Д. С. АНОНІМІЗАЦІЯ ДАНИХ В ІНТЕЛЕКТУАЛЬНИХ МЕДИЧНИХ СИСТЕМАХ. Радіоелектроніка та молодь у XXI столітті : Міжнар. молодіж. форум, м. Харків, 16–18 квіт. 2024 р. Харків, 2024.

14) Wong R., Fu C., Wang K., Pei J. Minimality attack in privacy preserving data publishing. *33rd International Conference on Very Large Data Bases (VLDB)*. 2007. P. 543–554.

15) LeFevre K., DeWitt D., Ramakrishnan R. Incognito: Efficient full-domain k-anonymity. *ACM SIGMOD International Conference on Management of Data (SIGMOD)*. 2005. P. 49–60.

16) A review of anonymization for healthcare data / I. E. Olatunji et al. *Big Data*. 2022. URL: https://doi.org/10.1089/big.2021.0169 (date of access: 21.04.2024).

17) GeMSyD: Generic Framework for Synthetic Data Generation. URL: https://www.researchgate.net/publication/377349269_GeMSyD_Generic_Framework_for_Synthetic_Data_Generation (date of access: 11.05.2024)

18) Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Bengio Y. Generative adversarial nets. *Advances in neural information processing systems*. 2014. P. 2672–2680.

19) Extracting and composing robust features with Denoising Autoencoders / Vincent P. et al. *Proceedings of the 25th International Conference on Machine Learning*. Helsinki, Finland, 2008.

20) Hinton G. E., Salakhutdinov R. R. Reducing the dimensionality of data with neural networks. *Science*. 2006. Vol. 313, No. 5786. P. 504–507.

21) TIOBE Index – TIOBE. TIOBE. URL: https://www.tiobe.com/tiobe-index/ (дата звернення: 16.05.2024).

22) medGAN Source Code: GitHub repository. URL: https://github.com/mp2893/medgan (дата звернення: 20.04.2024).

23) Kolmogorov–Smirnov test. Wikipedia. URL: https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test (date of access: 10.05.2024).

24) An automated technique for identifying associations between medications, laboratory results and problems / Wright A. et al. *J Biomed Inform.* 2010. Vol. 43, No. 6. P. 891–901.

25) Diagnostic analysis of patients with essential hypertension using association rule mining / Shin A. M. et al. *Healthc Inform Res*. 2010. Vol. 16, No. 2. P. 77–81.

26) Wu J., Roy J., Stewart W.F. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. *Med Care*. 2010. Vol. 48, No. 6. P. 106–113.

27) Vasa J., Thakkar A. Deep learning: differential privacy preservation in the era of big data. *Journal of Computer Information Systems*. 2022. P. 1–24. URL: https://doi.org/10.1080/08874417.2022.2089775 (date of access: 10.05.2024).

28) Baig Z., Mekala S. H., Zeadally S. Ransomware attacks of the COVID-19 pandemic: Novel strains, victims, and threat actors. *IT Professional*. 2023. No. 25(5). P. 37–44.

29) Apriori algorithm for the data mining of global cyberspace security issues for human participatory based on association rules / Z. Li et al. *Frontiers in Psychology*. 2021. Vol. 11. URL: https://doi.org/10.3389/fpsyg.2020.582480 (date of access: 10.05.2024).

30) Воронова Д. С. Методи штучного інтелекту для захисту медичних даних від гібридних загроз. *Гібридні війни сучасності: стійкість та протидія гібридним загрозам* : Всеукр. студент. науково-практ. конф. (м. Острог, 28 берез. 2024 р.). Острог, 2024.

ДОДАТОК А

Сертифікати про участь у конференціях з теми кваліфікаційної роботи

Рисунок А.1 – Сертифікат про участь у Всеукраїнській студентській науково-практичній конференції «Гібридні війни сучасності: стійкість та протидія гібридним загрозам»

Рисунок А.2 – Диплом про здобуття 3 місця в Міжнародному молодіжному форумі «Радіоелектроніка і молодь у XXI столітті»

ДОДАТОК Б

Відомість кваліфікаційної роботи

Позначення	Найменування	Дод. відомості
	Текстові документи	
1.	Пояснювальна записка	93 с.
	Інші документи	
2.	Презентаційні матеріали	__ плакатів

.	.	Прізвище та ініціали.	Підп.	Дата			
Розробив		Воронова Д.С.			Анонімізація великих даних в інтелектуальних медичних системах	Шифр групи	Код напр./спец.
Перевірив		Чала Л.Е.				СШм-22-3	122
Н.контр.		Малєєва І.А.				ХНУРЕ кафедра ШІ	
Затв.		Філатов В.О.					